EDUCATION**SECTOR** REPORTS

February 2009

BEYOND THE BUBBLE:

Technology and the Future of Student Assessment

By Bill Tucker

ACKNOWLEDGEMENTS

Thanks to my Education Sector colleagues Kevin Carey and Elena Silva for their help in thinking about the issues of assessment and technology, and to Thomas Toch and Andrew Rotherham for their support in the writing and editing of this paper. Thanks also to Abdul Kargbo for his support in designing graphics and formatting the report. Robin Smiles deserves special thanks for her careful editing and diligence in managing the production of this report. Research assistants Sumner Handy and Sara Yonker provided invaluable help with the research and ideas contained in this report. My sincere appreciation also goes to the many people who were kind enough to read and comment on an earlier draft of this paper, including Robin Chait, Randy Bennett, Drew Gitomer, Scott Marion, Michael Russell, David Coleman, Charles Barone, and Fritz Mosher. Finally, thank you to the dozens of educators, researchers, policy analysts, and experts who graciously offered their insights and knowledge to me throughout the research and writing of this report.

This publication was made possible by a grant from Carnegie Corporation of New York. The statements made and views expressed are solely the responsibility of the author.

ABOUT THE AUTHOR

BILL TUCKER is Chief Operating Officer of Education Sector. He can be reached at btucker@educationsector.org.

ABOUT EDUCATION SECTOR

Education Sector is an independent think tank that challenges conventional thinking in education policy. We are a nonprofit, nonpartisan organization committed to achieving measurable impact in education, both by improving existing reform initiatives and by developing new, innovative solutions to our nation's most pressing education problems.

1201 Connecticut Ave., N.W., Suite 850, Washington, D.C. 20036
202.552.2840 • www.educationsector.org

ABOUT THIS SERIES

This report is a product of Education Sector's Next Generation of Accountability initiative. The initiative seeks to strengthen public education by examining key elements of accountability, for instance, who should be responsible for student success and how they should be held responsible. Our work seeks to build on the strengths of current school accountability systems, more fully and effectively measure the depth and breadth of students' educational experiences, and encourage educators, parents, policymakers, and the larger public to pursue educational equity and excellence for all students.

Other reports in this series include *Measuring Skills for the 21st Century*, by Elena Silva.

Students today are growing up in a world overflowing with a variety of high-tech tools, from computers and video games to increasingly sophisticated mobile devices. And unlike adults, these students don't have to adjust to the information age—it will be all they've ever known. Their schools are gradually following suit, integrating a range of technologies both in and outside of the classroom for instructional use. But there's one day a year when laptops power down and students' mobile computing devices fall silent, a day when most schools across the country revert to an era when whiteboards were blackboards, and iPhones were just a twinkle in some techie's eye—testing day.

Since the IBM Type 805 Test Scoring Machine first hit the market in 1938, fill-in-the-bubble test score sheets and scanners have remained the dominant technologies used in local, state, and national assessments.[1] And underlying these pre-World War II technologies are approaches to testing from the same era. They rely heavily on multiple-choice question types and measure only a portion of the skills and knowledge outlined in state educational standards. They do not align well with what we know about how students learn. Nor do they tell us very much about how to help students do better. As a result, at a time when students are tested more than ever—and test results are used to make critical judgments about the performance of schools, teachers, and students—our testing methods don't serve our educational system nearly as well as they should.

States have slowly begun to adapt new technologies, such as the Internet, to student testing. Just over half the states, for instance, use computers to deliver a portion of the annual state testing programs mandated by the federal No Child Left Behind Act (NCLB).[2] But, for the most part, these states' investments in technology have not led to fundamental changes in our approaches to testing. Mostly, these investments have simply made old approaches to assessment more efficient. Even the most technologically advanced states have done little except replace the conventional paper-based, multiple-choice, fill-in-the-bubble tests with computerized versions of the same.[3] Overall, the types of skills tests measure, and what the test results can tell us, have remained essentially the same.

Technology, however, has the potential to do more than just make our current approach to testing more efficient. A growing number of testing and learning experts argue that technology can dramatically improve assessment—and teaching and learning. Several new research projects are demonstrating how information technology can both deepen and broaden assessment practices in elementary and secondary education, by assessing more comprehensively and by assessing new skills and concepts. All of which can strengthen both national standardized tests like the National Assessment of Educational Progress (NAEP) and classroom-based tests meant to help teachers improve their instruction.

These new technology-enabled assessments offer the potential to understand more than whether a student answered a test question right or wrong. Using multiple forms of media that allow for both visual and graphical representations, we can present complex, multi-step problems for students to solve, and we can collect detailed information about an individual student's approach to problem solving. This information may allow educators to better comprehend how students arrive at their answers and learn what those pathways reveal about students' grasp of underlying concepts, as well as to discover how they can alter their instruction to help move students forward. Most importantly, the new research projects have produced assessments that reflect what cognitive research tells us about how people learn, providing an opportunity to greatly strengthen the quality of instruction in the nation's classrooms. Other

fields, such as military training and medical education, are already using technology-enabled assessment to enhance teaching and learning.

But technology alone cannot transform assessment. Fundamentally changing our approach to testing in our public education system would not be easy. Logistical and funding challenges that often impede efforts to maintain, administer, and update schools' technological infrastructure would have to be overcome. New assessment models must not erode efforts to promote high expectations for all students; nor should they disadvantage low-income schools and students with currently limited access to technology. And new approaches to assessment would have to be aligned with standards, curricula, professional development, and instruction to be successful.

Still, the convergence of powerful new computer technologies and important new developments in cognitive science hold out the prospect of a new generation of student testing that could contribute to significant improvements in teaching and learning in the nation's classrooms.

A Decade of Incremental Progress

Educational researchers and testing experts from around the world have been writing about technology's potential to transform assessment for more than a decade.

The National Academy of Sciences, in its landmark 2001 report, *Knowing What Students Know: The Science and Design of Educational Assessment*, proclaimed it an "opportune time" to fundamentally rethink assessment, citing advances in technology, statistical modeling, and the sciences of thinking and learning.[4] New technology-enabled assessments, supported by research on how students learn, experts argued, would allow us to present complex, multi-step problems and record descriptive data about strategies used and actions taken by students. These data could then be used to adapt instruction by creating a better understanding about students' knowledge, and their conceptual understanding and cognitive development, which would lead not only to better assessment but to significant improvements in instruction and learning.

Similarly, in the late 1990s, Randy Bennett, a scientist at the Educational Testing Service (ETS) who directed the Technology-Based Assessment Project for the National Assessment of Educational Progress (NAEP), predicted that technology would enable educational testing to reinvent itself in three stages. First, technology would increase efficiency by automating existing processes. Secondly, test questions, formats for response, and scoring would become more sophisticated, allowing for the possibility of measuring new skills and measuring currently tested areas more comprehensively. At this stage, Bennett argued, technology would enable a new generation of simulations, and performance assessment would play an essential role. And, thirdly, Bennett envisioned testing merging with instruction, which would allow teachers and students to use feedback from testing to adjust teaching to improve student achievement.[5]

But in the main, the changes that Bennett and the National Academies envisioned have not taken place.

As researchers and cognitive scientists were beginning to recognize information technology's potential, the states were backing away from performance-based assessments, which were designed to mirror more complex, or real-world tasks. In the late 1980s and early 1990s, states began to experiment with using projects, portfolios, exhibitions, and other performance-based activities to measure content mastery.[6] The goal, writes Lorrie Shepard, dean of the University of Colorado's School of Education, was to produce assessments that "more faithfully reflect how learning would be used in non-test situations," assessments that were "guided by an underlying theory of teaching and learning drawn from the cognitive sciences."[7]

But the states' performance assessments were costly and technically inadequate for use in school accountability systems. A 1992 report published by the RAND Corporation on a portfolio assessment program in Vermont found significant problems with the reliability of the program's test scores.[8] It was difficult "to make scores comparable in meaning from year-to-year and from school-to-school," explains Harvard professor and measurement expert Daniel Koretz, who authored the report.[9] States, therefore, began to move away from performance-based assessment systems, back to less-expensive multiple-choice assessments. The demise of large-scale performance-based assessment systems also slowed efforts to link cognitive science with standards-based reform.

The enactment of NCLB in 2002 further complicated attempts to develop new types of testing. NCLB, which mandates that states give annual tests in reading and math in grades 3-8 and once in high school, resulted in a sizeable increase in the number of standardized tests given each year—now more than 45 million—creating a situation in which both test- and policymakers scrambled just to get the tests into the hands of teachers and students.[10] This tremendous increase in test taking, combined with the limited capacity of state departments of education and the nation's testing industry, encouraged state testing officials and testing companies to continue to use the same kinds of tests instead of pursuing innovations in assessment.

And, at key times, NCLB requirements, along with the relative immaturity of new assessment technologies, further slowed the development of new testing models. For instance, while NCLB does not prevent the use of computer-adaptive tests, which adjust the level of question difficulty based on students' answers to previous questions, it does require that tests align with state content standards and that each student be assessed at his or her official grade level.[11] (Lawmakers wanted to ensure that test results would be comparable from student to student and create common standards for all students, regardless of their backgrounds.) By 2002, Idaho and South Dakota were implementing such tests statewide to elementary and secondary students. But these early adaptive tests used by states adjusted to test low-achieving students' performance by presenting items that were below their grade level. Many of the tests were also plagued by technical and content issues.[12] As a result, the U.S. Department of Education would not allow these states to use computer-adaptive tests to meet NCLB requirements.

In 2007, the U.S. Department of Education approved Oregon's request to use a within-grade-level computer-adaptive test for NCLB-mandated state assessments.[13] But, in the meantime, many states signed multi-year contracts and spent millions of dollars investing in traditional, fixed-format tests.[14] The dot-com crash in the early 2000s and resulting state budget shortfalls also dampened enthusiasm for technology-based innovations in student assessment.

Despite this slow progress, assessment experts believe that testing will increasingly be delivered via computer and the Internet—especially as a way to continue to increase the efficiency of testing systems.[15]

Internet-based testing, for instance, eliminates the physical distribution, storage, and collection of test booklets and materials, along with data entry and scanning. Digital delivery and scoring saves time and accelerates the speed with which states can analyze and distribute test results. In 2008, fully 27 states delivered at least one of their state assessment tests via computer.[16] The most prominent, Virginia, administered more than 1.4 million tests online in the spring of 2008.[17]

Computer-adaptive tests offer a different type of efficiency. These tests can produce a more reliable estimate of student achievement using fewer items than required for a traditional test.[18] Since the test quickly adapts to a student's skill level, this form of testing eliminates the need for test items that don't yield information about a student's ability. Answers to easy questions, for example, offer little information to help assess a high-achieving student's specific level. Similarly, overly difficult questions provide little guidance on a low-achieving student's specific level. Whereas a one-size-fits-all test might need to employ 75 test questions to get enough data on students at various levels (for example, 25 questions at low, medium, and high levels), a computer-adaptive test could offer 50 questions instead. Moreover, since these questions are focused on the student's particular achievement level, the test can provide more specific evidence about that student's performance.[19] While computer-adaptive tests are only used for NCLB-mandated assessments in Oregon, they are increasingly used at the district level as practice and benchmark tests.[20]

The efficiencies gained from computer-based testing don't merely reduce the time and money used to administer testing programs. Incorporating automated essay scoring, a technology already in use on standardized tests such as the GMAT, the standardized test used for business-school admissions, enables assessments to test conceptual understanding and writing skills through open-ended, essay responses.[21] In addition, more efficient tests may make it possible to develop more flexible testing programs. Rather than yearly testing, portions of the test could be given throughout the year, offering a more accurate sample of students' progress over time.

For classroom or school-level assessments, results can be made available immediately to teachers, administrators, and district officials. They can also provide a greater connection to instruction, giving educators the chance

to collect information that can be used proactively to inform instruction, rather than only retroactively to gauge success. For example, automated essay scoring allows students to improve drafts with automated feedback. More periodic, flexible, and efficient testing will allow teachers to more easily embed assessment into current instructional processes.

Promising Models

At the same time, a number of promising research projects are beginning to explore the potential of technology to transform testing in more fundamental ways. They suggest that the technology-enabled assessment system that Bennett and others envisioned is indeed possible—a system that's both deeper and broader, able to test knowledge and skills more thoroughly and to test skills and concepts that haven't been measured in the past, and a system that reflects far more fully what we know about how students learn.

Testing Complex Skills

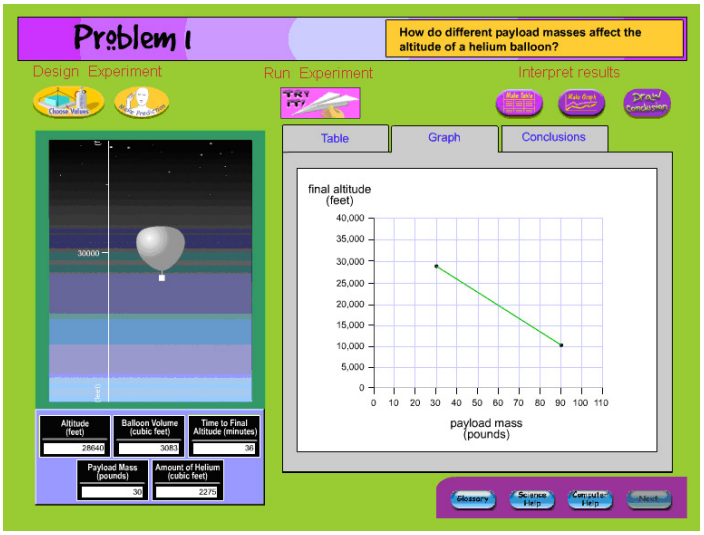
One of the largest efforts to pilot new forms of technology-based assessment is the Problem Solving in Technology-Rich Environments (TRE) project. It was launched in the spring of 2003, when a nationally representative sample of 2,000 students participated in a study to explore how information technology could be incorporated into the country's "gold standard" for assessment—the National Assessment of Educational Progress (NAEP). The goal was to create scenarios "that would feature … the kind of exploration characteristic of real-world problem solving."[22]

TRE tested scientific inquiry skills such as the ability to find information about a given topic, judge what information is relevant, plan and conduct experiments, monitor one's efforts, organize and interpret results, and communicate a coherent interpretation. In one component, eighth-graders used a simulated helium balloon to solve problems of increasing complexity about relationships among buoyancy, mass, and volume. For example, the students were asked to determine the relationship between payload mass and balloon altitude. To solve the problem, students gathered evidence by running simulated experiments using a variety of different payload masses. Once they had enough evidence, they submitted their conclusions using both open-ended and multiple-choice responses.[23]

TRE demonstrates several unique capabilities of technology-enabled assessments. First, technology permits the presentation of more complex, multi-step problems for students to solve. Multiple forms of media, such as the animated helium balloon and instrument panels in TRE, can present information in more useful and compelling ways than text alone. Technology-enabled assessments can present tasks based on complex data sets in ways that even elementary school students can use.[24] In TRE, for example, students see both visual and graphical representations showing what happens to the balloon during each experiment. (See Figure 1.)

Another example of technology-enabled assessment being used in science education is Floaters, a test given to students as part of the World Class Tests optional assessment program in the United Kingdom. The international initiative uses highly visual, engaging questions, enabling young students to be tested on an aspect of scientific method in a way not possible using paper and pencil. Students, for instance, use an interactive simulation to weigh a variety of foods, such as carrots, apples, and bananas, and observe whether the fruit floats in water. Students must then develop a hypothesis about the patterns they observe.[25]

Figure 1. TRE Exercise: The Relationship Between Payload Mass and Balloon Altitude



Source: Randy E. Bennett, Hilary Persky, Andrew R. Weiss, and Frank Jenkins, *Problem Solving in Technology-Rich Environments: A Report from the NAEP Technology-Based Assessment Project* (NCES 2007-466) (Washington, DC: National Center for Education Statistics, U.S. Department of Education, 2007). Retrieved on November 21, 2008 from http://nces.ed.gov/nationsreportcard/pubs/studies/2007466.asp.

Recording More Data

The problems in Floaters and TRE can be dynamic, presenting new information and challenges based on a student's actions. This allows students to take different approaches and even test multiple solutions. And critically, databases can record descriptive data about strategies used and actions taken by students. This provides a greater range of information, allowing instructors to make better judgments about student approaches, challenges, and performance.

In the TRE simulation exercise, for instance, every student action—such as which experiments they ran, which buttons they pushed, and what values they chose and in what order—is logged into a database. (See Figure 2.) Student actions, such as the quality of their experimental design choices, are evaluated using a set of rules and then scored based on statistical frameworks. These algorithms are linked across multiple skills, allowing students to be evaluated based on multiple points of evidence. And since each of the component skills can be traced back to observable student actions, instructors can gather detailed evidence to help determine why a student responded the way they did, helping to identify gaps in skill level, conceptual misunderstandings, or other information that could inform instruction.[26] Instead of just one data point, for example, a right or wrong answer, technology-enabled assessments can produce hundreds of data points about student actions and responses.

Linked to Classroom Instruction

Simulated exercises are particularly useful for assessing students' knowledge of interactions among multiple variables in a complex system, such as in an ecosystem. But, since these models assess both process and content, they require assessments that are closely linked with classroom instruction. This presents a problem for the broad use of these models. TRE, for example, restricted its assessment to scientific problem solving with technology—rather than science content—because NAEP cannot assume that students in the nation's some 14,000 school districts have all covered the same science content. Most of the time in science, however, as University of Maryland researcher Robert Mislevy explains, "it's not 'here's the situation in the world, and you give the answer.' Usually you have some hypotheses, some conjectures, but then you do something, and the world does something back. It's these cycles that really get at the nature of what model-based reasoning under constraints is really about."[27] But with large-scale tests such as NAEP, which Mislevy characterizes as "drop-in-from-the-sky" assessments, "you can't presume anything about what the examinees know about what they're going to be doing."[28]

In contrast, the Calipers project, funded by the National Science Foundation, seeks to develop high-quality, affordable performance assessments that can be used for both large-scale testing and in classrooms to inform instruction. Focused on physical science standards related to forces and motion, along with life sciences standards related to populations and ecosystems, Calipers engages students in problem-solving tasks such as determining the proper angle and speed to rescue an injured skier on an icy mountain. (See Figure 3.) Similar to TRE, Calipers captures descriptive data—describing the approach that a student took to solve the problem (choice of experimental values, formulas chosen), along with multiple-choice and open-ended responses. Calipers hopes to use these descriptive data, along with student reflection and self-assessment activities, to provide information to both students and teachers to guide learning and instruction.[29]

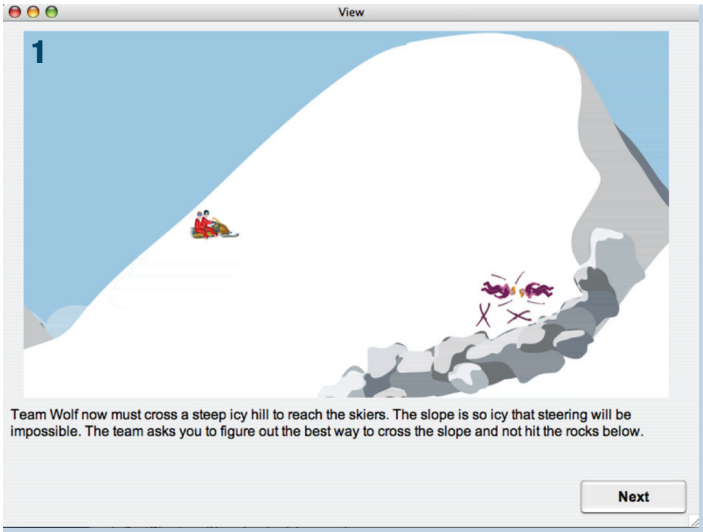
Figure 2. Logging One Eighth-Grader's Actions on a TRE Simulation Exercise (2003)

Time (in seconds)[1]	Action	Action choice
137	Begin problem 1	†
150	Choose values	90
155	Select mass	†
157	Try it	†
180	Make table	†
182	Selected table variables	Payload mass
185	Make graph	†
188	Vertical axis	Altitude
190	Horizontal axis	Helium

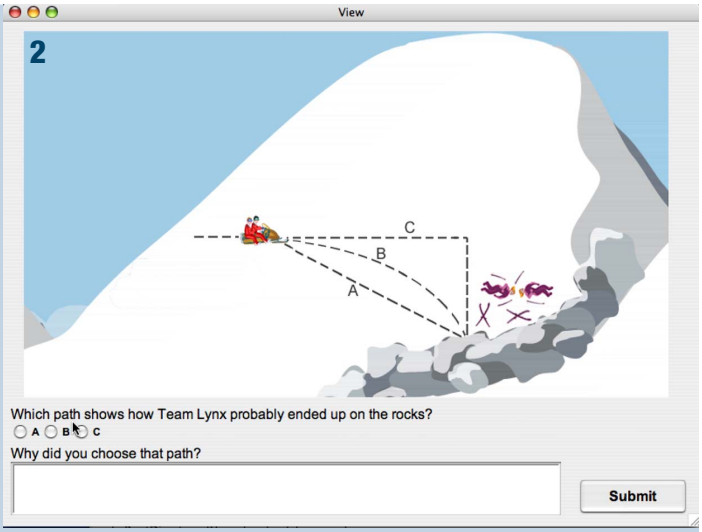
†Not applicable.
[1]These times include 137 seconds spent interacting with introductory material presented prior to problem 1.
Note: TRE=Technology-Rich Environments.

Source: Adapted from Table 2-1. Randy E. Bennett, Hilary Persky, Andrew R. Weiss, and Frank Jenkins, *Problem Solving in Technology-Rich Environments: A Report from the NAEP Technology-Based Assessment Project* (NCES 2007-466) (Washington, DC: National Center for Education Statistics, U.S. Department of Education, 2007). Retrieved on November 21, 2008 from http://nces.ed.gov/nationsreportcard/pubs/studies/2007466.asp.

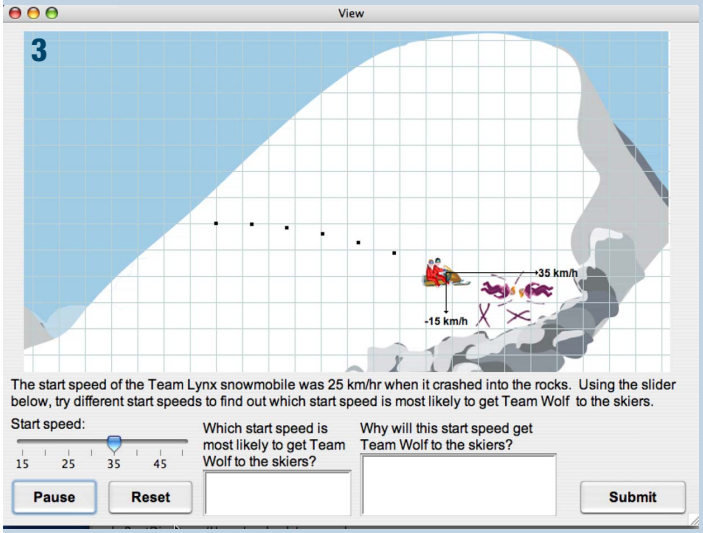
Table 3. CALIPERS Problem: Rescuing Injured Skiers



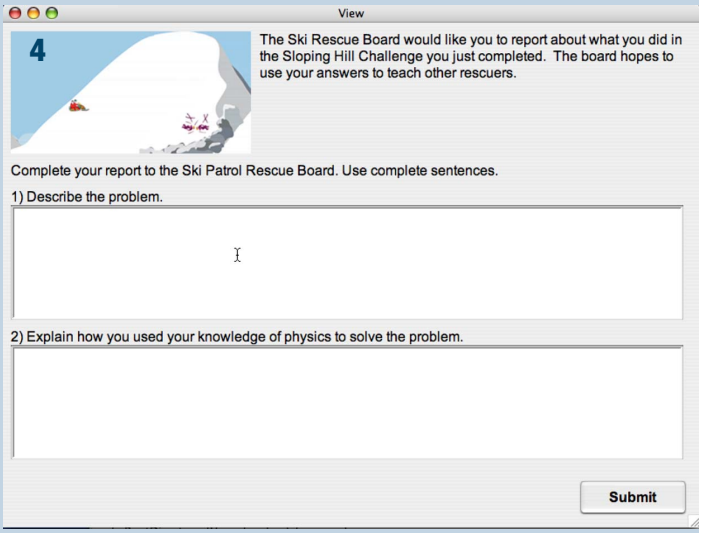
Test-takers are presented with a "real-life" problem that will test their understanding of physics principles.



Test-takers get a chance to choose from multiple options and explain their choice.



Test-takers can manipulate variables to achieve different outcomes, much as they would in the real world.



Test-takers are asked to demonstrate their understanding of the problem and how subject-matter knowledge helped them to solve it.

Source: http://calipers.sri.com/assessments.html.

Right Here in River City

Simulations also provide an opportunity to embed assessment into the learning process. The River City project, led by Harvard education professor Chris Dede, is a multi-user, virtual environment where middle-school students explore a mysterious illness in a turn-of-the-century town. Students learn by becoming scientists in River City's virtual world. With the project focused on inquiry practices, students make observations, "chat" with townspeople, develop hypotheses, and conduct experiments to determine the cause of the epidemic.[30]

Currently, River City uses traditional multiple-choice and teacher-graded assessments, such as a student-written report to the mayor outlining an action plan to eradicate the illness. But, in the future, researchers hope to use these traditional assessments in conjunction with the potential gold mine of descriptive data in the program's database. They are still working to develop algorithms to analyze and make use of the massive volumes of data River City captures about student actions. Jody Clarke, one of the River City researchers, says that the ultimate goal is to present data about what students are doing in the virtual environment in a way that helps teachers organize and

individualize instruction. She also believes that these data can be used to create performance-based summative assessments that are valid, reliable, and cost-effective.[31]

The Cisco Networking Academy, which teaches computer networking skills to 600,000 high school and college students each year in 160 countries around the world, provides another example of assessment that is embedded into learning. Functioning in over 9,000 different schools and 63 developing nations, such as Indonesia, Guinea, Mali, and El Salvador, the academy also demonstrates the potential for technology-enabled assessment at scale and in a wide variety of circumstances and settings.[32] A decade ago, employers complained that students graduating from the academy "do fine on the test, but you put them in front of a busted network and they have no idea what to do."[33] In response, the academy built Packet Tracer, a simulation and assessment engine that enables local instructors to create a variety of simulated computer network environments. With these simulations, students visualize how packets of data move across a network, further their understanding of how a network functions, and test their skills to identify and solve network problems.[34] Driven by a shared desire to assess how students perform in real-life situations, a number of other industries, such as architecture and accounting, are also using computer-based simulation for professional licensure.[35]

Perhaps even more importantly, the Cisco Networking Academy's technology, along with its integration with assessment, curriculum, and instruction, allows the academy to analyze data from tens of thousands of assessments to discover gaps and evaluate enhancements to instruction and curriculum at a program level.[36] Since it is much more defensible to make inferences from assessment data across larger numbers of students, these program-level uses of data are important potential benefits of technology-enabled assessments. Ideally, districts and states could use this type of information to rapidly test interventions across wide ranges of students, leading to the development of a powerful continuous improvement cycle.

Fully immersive simulations, such as those found in medical education and military training, point to further applications of technology. iStan, a life-like, sensor-filled mannequin that can talk, sweat, bleed, vomit, and have a heart attack, is used for medical-training purposes to simulate patient interactions and responses.[37] The U.S. Army has "instrumentalized" many of its war games and other performance exercises, using video and sensors to gather multiple sources of data about what is happening and when. As in the medical school simulations, these extensive data can illustrate multiple interactions among team members. This can lead to productive conversations about what happened, why, and ideas for improvement.[38] These types of assessments and simulated experiences are becoming more prevalent in higher education and the workplace. They engage participants in exercises to problem-solve realistic situations.

This focus on situated assessment, or assessing behavior in realistic situations, is increasingly important at a time when citizens and workers alike need to be able to communicate, collaborate, synthesize, and respond in flexible ways to new and challenging environments. Assessing the ability to approach new situations flexibly is challenging in our current paper-and-pencil environment.[39] "Real-life is not sequestered ….[what is important] is how you respond to feedback, not what you do in a feedback-free world," says John Bransford, University of Washington professor and a leading expert in cognition and learning technology.[40] Bransford is designing assessments that allow students to demonstrate not only what they can recall, but also how they can use their expertise. Technology-enhanced environments and virtual worlds, such as those found in medical training or River City, are necessary for students to practice and gain feedback in real-life, situated environments. In fact, Bransford notes, these efforts are "not possible without technology."

Aligning all the Parts

Education is a complex and decentralized public sector system, funded and governed at multiple levels. As a result, successful changes to assessment will require parallel, and equally challenging, revisions to standards, curriculum, instruction, and teacher training. Without deliberate attention from policymakers and educators in these areas, there is no guarantee that technology will fundamentally change core practices and methods in education, a field that has been notoriously impervious to change. Stanford University education historian Larry Cuban cautions that the "persistent dream of technology driving school and classroom changes has continually foundered in transforming teaching practices."[41] Just adding technology and hoping for educational transformation, without considering the content and practice of instruction, will do no more than automate existing processes, Cuban argues.

Standards and Cognitive Models

The cognitive research presented in the National Academy of Sciences 2001 report *Knowing What Students Know* stresses the importance of aligning assessments with curriculum and instruction and the need to base testing on a model of cognition and learning.[42] Yet, most state standards, as currently developed, focus on discrete sets of disconnected facts.[43] They do not provide a clear sense of where students are relative to desired goals, nor do they provide a complete road map for students and teachers to navigate.[44] In other words, our assessments do not align with what we know about how students learn and do not tell us enough about how to help students do better.

The disconnect is most evident in science education. Increasing global competition, a changing economy, and years of mediocre test results on international comparisons have sparked broad agreement among policymakers and educators that U.S. students must improve in the science, technology, engineering, and math (STEM) subject areas.[45] As such, many recognize that a different approach to teaching, learning, *and* assessment is needed. The National Academy of Sciences, in its 2007 report *Taking Science to School: Learning and Teaching Science in Grades K–8*, calls for "a redefinition of and a new framework for what it means to be proficient in science."[46] Their framework for science education, based on research on how students learn science, states that "content and process are inextricably linked in science." Scientific practices, such as inquiry, cannot be taught in isolation from the underlying concepts. But our tests, whether paper-based or online, focus almost exclusively on factual knowledge.

Also, while multi-user environments, simulations, and other technological domains offer many capabilities and opportunities, these tools are only as good as the cognitive models on which they are based. Mislevy, of the University of Maryland, cautions that "the evidentiary foundation … must be laid if we are to make sense of complex assessment data."[47] We can't use the data that these tools generate to inform assessment and instruction unless we have a greater understanding of how students learn within a domain. In a forthcoming article with fellow researcher Drew Gitomer, ETS' Bennett explains, "In principle, having a modern cognitive-scientific basis should help us build better assessments in the same way as having an understanding of physics helps engineers build better bridges."[48]

In fact, technology-enabled assessments expose the flaws in our current development of educational standards.[49] Most standards, for instance, are written as if we've asked teachers to ensure that their students can drive to a specific destination … let's say, Albuquerque. Our current assessments can tell us if a student has arrived, but don't tell us whether the students who haven't arrived are on their way, made a wrong turn, or have a flat tire. Technology-enabled assessments could in principle be like a GPS system, with the capability to frequently monitor and assess progress along the way. But a GPS is useless without the software that relates physical latitude and longitudinal coordinates back to a detailed map complete with roads, possible detours, and routes to Albuquerque. Similarly, to be transformative and to enhance teaching and learning, technology-enabled assessments will need to be dependent on a detailed understanding of how learning progresses in math, science, and various other disciplines. So far, however, our technological capabilities surpass our knowledge in these areas.[50]

For example, while there is much potential in the use of computer-based, multi-player gaming for both learning and assessment, we don't fully understand how gaming activities connect with the learning outcomes we are trying to teach and assess. Sigmund Tobias, a research scientist at Teachers College, Columbia University and J.D. Fletcher, a senior researcher at the Institute for Defense Analyses, in their review of research on gaming and learning, argue that even though a game may have similarities to what is being taught or assessed for real-life use, the important learning outcomes don't necessarily transfer. It's essential to analyze the actual cognitive tasks involved in the game and map them to the goals, they write.[51]

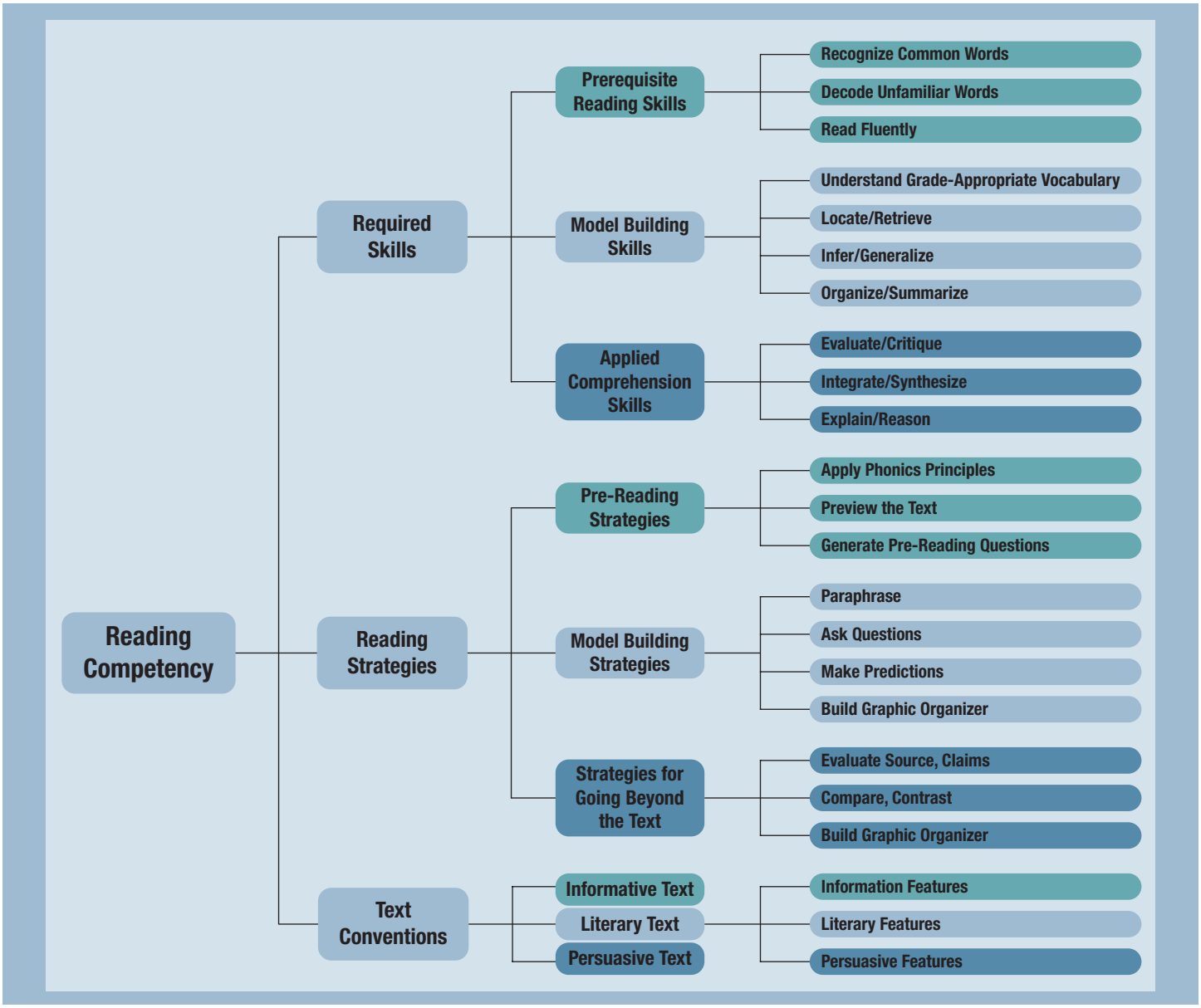
The Educational Testing Service's Cognitively Based Assessment of, for and as Learning (CBAL) research project provides an example of where both technology and the map come together. While the project is dependent on technology—it uses automated essay scoring and the research takes place in Portland, Maine, in schools with one-to-one laptop programs—the extensive research and development of a cognitive model for how students read and develop reading skills is also essential.

The cognitive model forms the bridge between two different uses of testing. Summative assessment describes what has been learned over time and is used to judge the performance of students or schools,

while formative assessment is meant to guide teaching and learning as part of the instructional process.[52] Projects built on cognitive models, such as CBAL and Calipers, attempt to build both summative and formative components, held together by a common conception of how students learn a particular subject.

For example, in CBAL, the model for reading competency includes three broad categories of required skills: pre-requisite reading skills, model building skills, and applied comprehension skills.[53] (See Figure 4.) Each of these categories is fully developed in the model and assessed during the periodic tests given over the course of a school year. Questions test applied literacy, with tasks such as evaluating text content for bias and evidence to support claims using a wide variety of sources, such as newspaper articles, encyclopedia entries, and diagrams. The cognitive model underlying CBAL ensures that the

Figure 4. Cognitive Model for Reading Assessment



Source: K. M. Sheehan and T. O'Reilly, "The Case for Scenario-Based Assessments of Reading Competency" (paper presented at the Assessing Reading in The 21st Century Conference: Aligning and Applying Advances in the Reading and Measurement Sciences, Philadelphia, PA., April 2008). Reprinted by permission of Educational Testing Service, the copyright owner. No endorsement of this publication by Educational Testing Service should be inferred.

project's summative assessments, meant to be used for accountability purposes, accurately align with and assess all of the various dimensions of reading. Still in its early stages, this conceptual model has also allowed CBAL researchers to begin differentiating students and instructional responses to those students based on their performance on the CBAL assessments.[54]

Similarly, research in the United Kingdom is testing how technology-enabled assessments, combined with advanced statistical and cognitive models, allow teachers to identify groups of readers with different patterns of performance even though the students' raw test scores may be similar. Teachers can then tailor instruction to four types of readers—reluctant readers, developing readers, reasoning readers, and involved readers.[55]

But without a sound evidentiary model and conceptual underpinning, technology-enabled assessment tools are just more efficient, faster, and accessible versions of the same old tests. For example, although many Internet-based benchmark tests are marketed as formative assessment products, most do not provide the specific information needed to improve instruction. Increasingly popular at the district level, these online test systems come equipped with banks of test questions for classroom teachers to use. However, rather than drawing from a common cognitive model, they are essentially just online variations of the types of questions found on state assessments. They provide practice on these question types and predict what a student will score on a state assessment test, but they don't tell a teacher *why* a student scored a certain way or how to modify instruction.

Effective Instruction

The most useful classroom-level applications of technology-enabled assessments, such as CBAL, River City, and others, provide descriptive data that give teachers better information about how students are progressing and why they are performing at their current levels. They also provide insight into what changes to instruction might be the most effective.

These formative uses of technology-enabled assessments are both promising and challenging. Formative assessment, when implemented effectively, is one of the few reforms to show significant effects on student gains—gains that research shows to be especially pronounced for the most struggling students.[56]

But this classroom-level use is highly dependent on effective teaching. Teachers must have deep content expertise and specific skills to understand and use the insights these sophisticated assessments produce. The teachers must also be open to an entirely new approach to instruction. David Niemi, formerly with UCLA's National Center for Research on Evaluation, Standards, and Student Testing, writes, "Teachers must achieve not only deep understanding of the subjects they teach but also broad knowledge about how that understanding typically develops over time and how it may be assessed."[57]

Andrew Boyle, a leading researcher on technology-enabled assessment (known as "e-assessment" in the United Kingdom), cautions that "at the present time, sophisticated tasks for e-assessment may be characterized as expensive and slow to develop, and not easily written by a non-specialist teacher."[58] Technology has enabled mass customization in a number of areas, but the challenge of creating high-quality assessments and simultaneously making them adaptable by teachers for easy classroom use is considerable.

Infrastructure and Equity

In addition to the teacher-quality challenge, most school technology infrastructures are still insufficient for advanced, large-scale, Internet-based testing. In *Computer-Based Assessment: Can It Deliver on Its Promise*, a 2001 policy brief published by WestEd, a nonprofit research agency, the authors concluded, "It seems only a matter of time before computer-based assessment defines the test administration process."[59] Yet, in a November 2008 feature on computer-based assessment in *Education Week*, the writer noted that "progress toward that goal has been slow, expensive, and fraught with logistical challenges."[60] Inadequate computer hardware, bandwidth constraints, and limited capacity to maintain, update, and administer schools' technological infrastructure have proven to be serious impediments to deploying online testing, especially for more secure, high-stakes state NCLB testing.

The serious challenges encountered in 2008 by NAEP in its initial testing of computer-based questions for the 2009 Science Assessment provide an instructive case study.

Initially, NAEP administrators relied on schools' computers and technology infrastructure to administer the student exams. However, due to a wide variety of problems with the schools' hardware, software configurations, and Internet access, technical problems impeded half the students in the first few weeks of the trial. Finally, NAEP resorted to bringing its own laptops into each school and equipped every student with a secure portable flash drive to store data. This experience, along with the tremendous logistical burdens and other costs associated with transporting and setting up laptops in each of the schools, led NAEP to significantly scale back its plans for technology-enabled testing in 2009. Interactive, technology-enabled questions will only be given to a small subset of schools participating in the 2009 NAEP.[61]

Moreover, concerns that technology-enabled assessment will disadvantage primarily low-income students that may have limited access to computers and other forms of technology must also be overcome. For example, if an assessment that is developed to assess math skills inadvertently tests computer skills, then the results will be biased against students with less computer access. Results from earlier NAEP technology-enabled assessment projects, Math Online and Writing Online, found that students with greater exposure to computers outside of school scored higher.[62] Other studies show no difference between paper and computer-based tests, indicating that more research is needed to understand how to avoid bias against students without computer familiarity.[63] At the same time, technology can help address the needs of specific populations, such as English-language learners or students with disabilities. (See sidebar, "The Benefits of Universal Design," page 12.)

Changing Course

Overcoming these barriers will be challenging. But so is the goal of helping all students reach challenging standards for learning and be prepared for future success. To reach these goals educators must be able to adapt instruction to account for a multitude of variables. This adaptive instruction is not possible without a deeper understanding of and better data about how students learn.[64]

Now with the impending reauthorization of NCLB, there's an opportunity to begin to chart a different course for the future of educational assessment, one that would also enhance teaching and learning, which is essential to meeting NCLB's goals. In fact, by requiring state officials to rapidly expand testing in American classrooms, NCLB helped bring the inadequacy of our current testing practices to the forefront and created a rare consensus among proponents, critics, teachers and policymakers—none are satisfied with the state of testing today.

With technology changing at a rapid pace, we have many of the tools to create vastly improved assessment systems and practices. Dozens of research projects, both here in the United States and in countries around the world, are beginning to demonstrate how technology, used in concert with what cognitive-scientific research tells us about how people learn, can improve formative assessment and enhance teaching and learning. And, the experiences from a variety of industries, ranging from medicine to accounting, shows that simulations and more advanced, technology-enabled assessments can also be used for large-scale and consequential testing.

And despite the infrastructure challenges, NAEP, which is perhaps our nation's most important large-scale test, is moving quickly to incorporate technology into its main assessment program. With a main key goal to "go beyond what can be measured using paper-and-pencil," the 2009 NAEP Science Assessment will use technology-based assessment tasks for a subsample of the students taking the test. In 2011 NAEP plans to go even further—that year's eighth-grade writing assessment will be delivered entirely by computer.[65]

Yet, because changes in assessment impact our entire educational system and infrastructure, from state agencies to test-makers to federal officials to classroom teachers, we won't see the real benefits from technology-enabled assessments—improved teaching and learning—without careful attention from policymakers and deliberate strategies to create change. It will take time to further research, build, and implement on a wide-scale the types of technology-enabled assessments described in this report. But in the meantime there are steps that policymakers, educators, and a variety of stakeholders can take immediately to ensure that progress moves much more quickly in the next decade.

To begin with, if we want to see dramatically better assessments, we can't continue to invest solely in the

The Benefits of Universal Design

On television sets all over the country, closed-captioning decoder chips allow exercisers on treadmills, fans at noisy sports bars, students learning a second language, and hearing-impaired persons to follow the dialogue on their favorite television shows. But decoder chips weren't always built into every television. Prior to built-in chips, hearing-impaired persons had to seek special accommodations, such as expensive set-top decoder boxes, to access closed captioning. Integrating decoder chips into every television not only offered much greater access to the hearing-impaired, but was highly cost-effective and provided valuable benefits to a wide range of television viewers.

Universal design, the concept behind this innovation, allows designers to address the divergent needs of special populations and at the same time, increase usability for everyone.[a] Experience from a wide variety of fields, from architecture to civil engineering to consumer products, shows that the application of universal design does not only provide wide benefits, but can save money by preventing costly retrofits and special accommodations.

Technology allows developers to apply these same universal design concepts to educational assessment. Digital materials are inherently flexible, making it feasible to customize materials and methods to each individual.[b] Rather than provide special accommodations, such as a separately printed test booklet with enlarged print for students with reduced vision, technology-enabled testing can embed a variety of accommodations into the same program. In an *Education Week* commentary, Andrew Zucker, author of *Transforming Schools With Technology* and a senior research scientist at the Concord Consortium, writes that "Computers enlarge typographical fonts, translate to and from English, convert text to speech, correct mistakes, and help teachers individualize instruction."[c] Applying universal design to assessment offers the potential to allow all students to benefit from a testing environment that adapts to meet individual needs.[d]

Other, more targeted accommodations are also possible. Researchers are investigating the use of graphical interfaces to make tests more accessible and valid for specific populations. It is difficult, for instance, to test English-language learners, students with low English proficiency, in subjects such as science without mistakenly testing their language skills instead. Rebecca Kopriva, a University of Wisconsin researcher, is currently testing computer-based science assessments that use animations and non-verbal communications to test concepts such as what happens to water molecules when they are heated.[e]

But, to be effective, these accommodations must go beyond just assessment. Accommodations built into technology-enabled assessments must also be available for students throughout the entirety of their educational experiences. In fact, the National Center for the Improvement of Educational Assessment notes that "If the first time a student sees a certain type of accommodation—especially if it is somewhat novel—is on the large-scale assessment, that accommodation will likely hinder instead of improve access."[f] As in every other aspect of assessment design, usability enhancements to assessment must also align with similar enhancements to curriculum and instruction.

[a] David H. Rose, Anne Meyer, Nicole Strangman, and Gabrielle Rappolt, *Teaching Every Student in the Digital Age* (Alexandria, VA: Association for Supervision and Curriculum Development, 2002). Retrieved December 10, 2008 from http://www.ascd.org/publications/books/101042.aspx.

[b] National Center on Accessing the General Curriculum, "Virtual Reality and Computer Simulations and the Implications for UDL Implementation: Curriculum Enhancements Report." Retrieved December 10, 2008, from http://www.k8accesscenter.org/training_resources/udl/documents/VirtualRealityUDL_000.pdf.

[c] Andrew Zucker, "Commentary: Smart Thinking About Educational Technology," *Education Week*, April 2, 2008.

[d] G. Madaus, M. Russell, and J. Higgins, *The Paradoxes of High-Stakes Testing: How They Affect Students, Their Parents, Teachers, Principals, Schools, and Society*.

[e] "Computer Simulations: Four Approaches to Interactive Student Assessment" (presentation at the Council of Chief State School Officers National Conference on Student Assessment, June 17, 2008).

[f] Marianne Perie and Scott Marion, "Computer Based Testing in Utah: A Summary of Key Issues and Stakeholder Input," an unpublished report from the Center for Assessment, February 29, 2008.

current practice. In addition to today's current federal dollars for assessment, federal policymakers should create a second, smaller pool of funding to support the research and development of the next generation of assessment technology and practice. This money should be focused not just on isolated research projects, but on applied applications that involve competitive awards and partnerships among researchers, psychometricians, testing companies, state officials, and educators. Rather than just award contracts, funds should support both early developmental work and provide large incentives for actual district or state-level implementation.

Also, because the cognitive and data models underlying technology-enabled assessments will be used to gauge student progress and guide instruction as much as possible and much more than today, they must be open for public review and improvement, ensuring that evaluators can test and enhance these models along the way. And, within the broad parameters of federal policy and coupled with rigorous evaluation, schools and educators participating in these innovative initiatives should have the freedom to use, as well as incentives for the use of, new assessments and assessment practices—both summative and formative. Given the importance of

science, the deficiency of current science assessment practices, and the number of promising research projects in this field, science education is a logical place to start.

We should plot a different course—one that maintains accountability goals but creates space for significant innovation and prioritizes the use of technology-enabled assessments not just for automation, but for substantive improvements in student achievement.

Endnotes

[1] Retrieved on September 25, 2008, from http://www-03.ibm.com/ibm/history/exhibits/specialprod1/specialprod1_9.html

[2] C.V. Bausell, "Tracking U.S. Trends," *Education Week*, March 27, 2008. For more on states' efforts to implement online testing, please see Katie Ash, "States Slow to Embrace Online Testing," *Education Week*, November 19, 2008.

[3] G. Madaus, M. Russell, and J. Higgins, *The Paradoxes of High-Stakes Testing: How They Affect Students, Their Parents, Teachers, Principals, Schools, and Society* (Charlotte, NC: Information Age Publishing. Forthcoming).

[4] James W. Pellegrino, Naomi Chudowsky, and Robert Glaser, eds., *Knowing What Students Know: The Science and Design of Educational Assessment*, Committee on the Foundations of Assessment, Board on Testing and Assessment, Center for Education, National Research Council (Washington, DC: The National Academies Press, 2001).

[5] Randy Bennett, *Reinventing Assessment: Speculations on the Future of Large-Scale Educational Testing* (Princeton, NJ: Educational Testing Service, 1998).

[6] Daniel Koretz, *Measuring Up: What Educational Testing Really Tells Us* (Cambridge, MA: Harvard University Press, 2008).

[7] Lorrie Shepard, "A Brief History of Accountability Testing: 1965-2007," in The Future of Test-Based Educational Accountability, eds. Katherine E. Ryan and Lorrie A. Shepard (New York: Routledge, 2008).

[8] Daniel Koretz, Daniel F. McCaffrey, Stephen P. Klein, Robert M. Bell, Brian M. Stecher, "The Reliability of Scores from the 1992 Vermont Portfolio Assessment Program" (Santa Monica, CA: RAND Corporation, 1992). Retrieved on September 22 from http://www.rand.org/pubs/drafts/DRU159/

[9] Daniel Koretz, *Measuring Up: What Educational Testing Really Tells Us*.

[10] See Thomas Toch, *Margins of Error: The Education Testing Industry in the No Child Left Behind Era* (Washington, DC: Education Sector, 2006) for more discussion of the effects of the rapid increase in standardized testing.

[11] Rhea R. Borja, "South Dakota Drops Online 'Adaptive' Testing," *Education Week*, January 29, 2003.

[12] Ibid.

[13] As of December 2008, Oregon remains the only state approved to use computer adaptive testing for NCLB assessments. In a letter to the Utah State Superintendent of Public Instruction, dated November 17, 2008, Assistant Secretary of Education Kerri L. Briggs denied a waiver for a pilot computer adaptive program in Utah. "It is vital that all students are validly and reliably tested on assessments aligned to a state's challenging academic content standards in order to obtain an accurate measure of student performance." The state "has not demonstrated that the pilot assessments comply with ESEA assessment requirements, including those for technical quality."

[14] G. Madaus, M. Russell, and J. Higgins, *The Paradoxes of High-Stakes Testing: How They Affect Students, Their Parents, Teachers, Principals, Schools, and Society*.

[15] Katie Ash, "States Slow to Embrace Online Testing," *Education Week*, November 19, 2008.

[16] C.V. Bausell, "Tracking U.S. Trends." *Education Week*, March 27, 2008.

[17] "eTesting at the State Level—Lessons Learned" (presentation at the Council of Chief State School Officers conference, June 16, 2008).

[18] G. Madaus, M. Russell, and J. Higgins, *The Paradoxes of High-Stakes Testing: How They Affect Students, Their Parents, Teachers, Principals, Schools, and Society*.

[19] Computer-adaptive tests are most efficient when used for selection or as a gauge of overall ability, for instance, on the GRE. They do not provide as many scoring efficiencies in a standards-based environment if educators are attempting to understand how well a student did in each particular aspect of a subject, for instance, math, because then the test needs to ask questions in each specific aspect. Researchers continue to caution that there are still limitations to computer adaptive testing, such as concerns about whether scores are well-estimated for all examinees and the need for large item pools to ensure test security.

[20] For more information on computer-adaptive testing, please see Katie Ash, "Adjusting to Test Takers," *Education Week,* November 19, 2008.

[21] On the GMAT, a human reader also scores each essay, and additional reviews are conducted if the scores between the computer and human raters disagree significantly. See Greg Miller, "Computers as Writing Instructors," *SCIENCE*, January 2, 2009. Findings across a number of studies consistently show comparability of human and computer scoring at levels approximating the agreement between two human scorers. See Edys Quellmalz and James Pellegrino, "Technology in Testing," *SCIENCE*, January 2, 2009.

[22] Randy E. Bennett, Hilary Persky, Andrew R. Weiss, and Frank Jenkins, *Problem Solving in Technology-Rich Environments: A Report from the NAEP Technology-Based Assessment Project* (NCES 2007-466) (Washington, DC: National Center for Education Statistics, U.S. Department of Education, 2007). Retrieved on November 21, 2008 from http://nces.ed.gov/nationsreportcard/pubs/studies/2007466.asp.

[23] Ibid.

[24] Jim Ridgway and Sean McCusker, "Using Computers to Assess New Educational Goals," *Assessment in Education* 10, no. 3 (November 2003).

[25] Andrew Boyle, "Sophisticated Tasks in E-Assessment: What Are They and What Are Their Benefits?" (paper presented at 2005 International Computer Assisted Assessment Conference, Loughborough University, United Kingdom). Retrieved on July 17, 2008 from http://www.caaconference.com/pastConferences/2005/proceedings/BoyleA2.pdf

[26] Randy E. Bennett, Hilary Persky, Andrew R. Weiss, and Frank Jenkins, *Problem Solving in Technology-Rich Environments: A Report from the NAEP Technology-Based Assessment Project*.

[27] Robert Mislevy, University of Maryland, personal communication, February 1, 2008.

[28] Ibid.

[29] Quellmalz, Edys S., et al. "Exploring the Role of Technology-Based Simulations in Science Assessment: The Calipers Project" in *Assessing Science Learning: Perspectives From Research and Practice*, eds. Janet Coffey, Rowena Douglas, and Carole Stearns (Arlington, VA: National Science Teachers Association Press, 2008); "Computer Simulations: Four Approaches to Interactive Student Assessments" (presentation at the Council of Chief State School Officers conference, June 17, 2008).

[30] Christopher Dede, "Transforming Education for the 21st Century: New Pedagogies that Help All Students Attain Sophisticated Learning Outcomes," Commissioned by the NCSU Friday Institute, February, 2007. Retrieved from http://www.gse.harvard.edu/~dedech/Dede_21stC-skills_semi-final.pdf

[31] Jody Clarke, Harvard Graduate School of Education, personal communication, March 14, 2008.

[32] Statistics provided by Cisco Networking Academy, retrieved from http://www.cisco.com/web/learning/netacad/academy/index.html on December 5, 2008.

[33] Robert Mislevy, University of Maryland, personal communication, February 1, 2008.

[34] John Behrens, director, Networking Academy Learning Systems Development, Cisco Systems, Inc., personal communication, April 2, 2008.

[35] See, for example, the computer-based simulations used in the United States Medical Licensing Examination Step III test. Other examples include architecture (NCARB) and accountancy (AICPA); Randy Bennett, Educational Testing Service, personal communication, November 21, 2008.

[36] John Behrens, personal communication.

[37] Personal demonstration, Games, Learning, and Society Conference, Madison, Wis., July 11, 2008; "Remote-controlled 'man' Called IStan Trains Healthcare Professionals," *Medical News Today*, July 28, 2008. Retrieved on August 20, 2008, http://www.medicalnewstoday.com/articles/116286.php

[38] Dexter Fletcher, Institute for Defense Analyses, personal communication, July 09, 2008.

[39] John D. Bransford, Ann L. Brown, and Rodney R. Cocking, eds., *How People Learn: Brain, Mind, Experience, and School*, Committee on Developments in the Science of Learning, National Research Council (Washington, DC: The National Academies Press, 1999).

[40] John Bransford, personal communication, March 19, 2008.

[41] Larry Cuban, "Techno-Reformers and Classroom Teachers," *Education Week*, October 9, 1996.

[42] James W. Pellegrino, Naomi Chudowsky, and Robert Glaser, eds., *Knowing What Students Know: The Science and Design of Educational Assessment*, Committee on the Foundations of Assessment, Board on Testing and Assessment, Center for Education, National Research Council (Washington, DC: The National Academies Press, 2001).

[43] Richard A. Duschl, Heidi A. Schweingruber, and Andrew W. Shouse, eds., *Taking Science to School: Learning and Teaching Science in Grades K–8*, Committee on Science Learning, Kindergarten through Eighth Grade, (Washington, DC: National Academy of Sciences, 2007).

[44] Margaret Heritage, "Learning Progressions: Supporting Instruction and Formative Assessment," (paper prepared for the Council of Chief State School Officers, 2008).

[45] See "Technology Counts: The Push to Improve Science, Technology, Engineering, and Mathematics," *Education Week*, March 27, 2008, for more background on calls for improved science education,

[46] Richard A. Duschl, Heidi A. Schweingruber, and Andrew W. Shouse, eds., *Taking Science to School: Learning and Teaching Science in Grades K–8*.

[47] Robert J. Mislevy, "Leverage Points for Improving Educational Assessment," U.S. Department of Education Office of Educational Research and Improvement Award #R305B60002, National Center for Research on Evaluation, Standards, and Student Testing (CRESST) Center for the Study of Evaluation (CSE) Graduate School of education and Information Studies, University of California, Los Angeles, 2000.

[48] Randy E. Bennett and Drew H. Gitomer, "Transforming K–12 Assessment," in *Assessment Issues of the 21st Century*, eds. C. Wyatt-Smith and J. Cumming (New York: Springer Publishing Company, Forthcoming).

[49] Chris Brown, senior vice president, Pearson Education, personal communication, March 12, 2008.

[50] Derek Briggs, professor, University of Colorado, personal communication, April 18, 2008.

[51] Sigmund Tobias and J.D. Fletcher, "What Research Has to Say About Designing Computer Games for Learning," *Educational Technology* 47 (2007).

[52] Definitions taken from "Building Balanced Assessment Systems to Guide Educational Improvement" Doris Redfield, Ed Roeber, and Rick Stiggins, background paper for the National Conference on Student Assessment, June 15, 2008.

[53] Tenaha O'Reilly and Kathleen M. Sheehan, *Cognitively Based Assessment of, for, and as Learning: A 21st Century Approach for Assessing Reading Competency* (Princeton, NJ: Educational Testing Service, 2008). Retrieved November 21, 2008, from http://www.cogsci.rpi.edu/csjarchive/proceedings/2008/pdfs/p1613.pdf

[54] Drew Gitomer, researcher, Educational Testing Service, personal communication, March 11, 2008; also see ibid above.

[55] Chris Whetton and Marian Sainsbury, National Foundation for Educational Research, UK, "E-Assessment for Improving Learning" (paper presented at 33rd International Association for Educational Assessment, Baku, Azerbaijan, September 2007).

[56] Dylan Wiliam, "Changing Classroom Practice,", *Educational Leadership* 65, no. 4 (Dec./Jan. 2008).

[57] David H. Niemi, "Cognitive Science, Expert-Novice Research, and Performance Assessment," *Theory into Practice* 36, no. 4, New Directions in Student Assessment, (Autumn 1997).

[58] Andrew Boyle, "Sophisticated Tasks in E-Assessment: What Are They and What Are Their Benefits?"

[59] Stanley Rabinowitz and Tamara Brandt, Computer-based Assessment: Can it Deliver on its Promise? (San Francisco, CA: WestEd, 2001).

[60] Katie Ash, "States Slow to Embrace Online Testing."; see also Andrew Trotter, "Online Testing Demands Careful Planning," *Education Week Digital Directions*, June 20, 2007, for further discussions of challenges faced by states implementing online testing, including Virginia.

[61] "At Last! Computer Based Testing for Main NAEP," (presentation at the Council of Chief State School Officers National Conference on Student Assessment, June 17, 2008).

[62] Randy E. Bennett, J. Braswell, A. Oranje, B. Sandene, B. Kaplan, and F. Yan, "Does It Matter If I Take My Mathematics Test on Computer? A Second Empirical Study of Mode Effects in NAEP," *Journal of Technology, Learning, and Assessment* 6, no. 9 (2008). Retrieved November 10, 2008 from www.jtla.org. Also, N. Horkay, R. E. Bennett, N. Allen, B. Kaplan, and F. Yan, "Does It matter If I Take My Writing Test on Computer? An Empirical Study of Mode Effects in NAEP," *Journal of Technology, Learning and Assessment* 5, no. 2 (2006). Retrieved November 21, 2008 from http://escholarship.bc.edu/jtla/vol5/2/

[63] See Vol. 6 of *The Journal of Technology, Learning and Assessment* (JTLA), www.jtla.org, for additional studies and more discussion of this issue.

[64] For more discussion of assessment's role in adapting instruction, see Thomas B. Corcoran and Frederic A. (Fritz) Mosher, *The Role of Assessment in Improving Instruction* (Philadelphia, PA: Consortium for Policy Research in Education (CPRE) Center on Continuous Instructional Improvement (CCII), October, 2007).

[65] "At Last! Computer Based Testing for Main NAEP," (presentation at the Council of Chief State School Officers conference, June 17, 2008).

EDUCATION**SECTOR** REPORTS

February 2009

BEYOND THE BUBBLE:

Technology and the Future of Student Assessment

By Bill Tucker

ACKNOWLEDGEMENTS

Thanks to my Education Sector colleagues Kevin Carey and Elena Silva for their help in thinking about the issues of assessment and technology, and to Thomas Toch and Andrew Rotherham for their support in the writing and editing of this paper. Thanks also to Abdul Kargbo for his support in designing graphics and formatting the report. Robin Smiles deserves special thanks for her careful editing and diligence in managing the production of this report. Research assistants Sumner Handy and Sara Yonker provided invaluable help with the research and ideas contained in this report. My sincere appreciation also goes to the many people who were kind enough to read and comment on an earlier draft of this paper, including Robin Chait, Randy Bennett, Drew Gitomer, Scott Marion, Michael Russell, David Coleman, Charles Barone, and Fritz Mosher. Finally, thank you to the dozens of educators, researchers, policy analysts, and experts who graciously offered their insights and knowledge to me throughout the research and writing of this report.

This publication was made possible by a grant from Carnegie Corporation of New York. The statements made and views expressed are solely the responsibility of the author.

ABOUT THE AUTHOR

BILL TUCKER is Chief Operating Officer of Education Sector. He can be reached at btucker@educationsector.org.

ABOUT EDUCATION SECTOR

Education Sector is an independent think tank that challenges conventional thinking in education policy. We are a nonprofit, nonpartisan organization committed to achieving measurable impact in education, both by improving existing reform initiatives and by developing new, innovative solutions to our nation's most pressing education problems.

1201 Connecticut Ave., N.W., Suite 850, Washington, D.C. 20036
202.552.2840 • www.educationsector.org

ABOUT THIS SERIES

This report is a product of Education Sector's Next Generation of Accountability initiative. The initiative seeks to strengthen public education by examining key elements of accountability, for instance, who should be responsible for student success and how they should be held responsible. Our work seeks to build on the strengths of current school accountability systems, more fully and effectively measure the depth and breadth of students' educational experiences, and encourage educators, parents, policymakers, and the larger public to pursue educational equity and excellence for all students.

Other reports in this series include *Measuring Skills for the 21st Century*, by Elena Silva.

Students today are growing up in a world overflowing with a variety of high-tech tools, from computers and video games to increasingly sophisticated mobile devices. And unlike adults, these students don't have to adjust to the information age—it will be all they've ever known. Their schools are gradually following suit, integrating a range of technologies both in and outside of the classroom for instructional use. But there's one day a year when laptops power down and students' mobile computing devices fall silent, a day when most schools across the country revert to an era when whiteboards were blackboards, and iPhones were just a twinkle in some techie's eye—testing day.

Since the IBM Type 805 Test Scoring Machine first hit the market in 1938, fill-in-the-bubble test score sheets and scanners have remained the dominant technologies used in local, state, and national assessments.[1] And underlying these pre-World War II technologies are approaches to testing from the same era. They rely heavily on multiple-choice question types and measure only a portion of the skills and knowledge outlined in state educational standards. They do not align well with what we know about how students learn. Nor do they tell us very much about how to help students do better. As a result, at a time when students are tested more than ever—and test results are used to make critical judgments about the performance of schools, teachers, and students—our testing methods don't serve our educational system nearly as well as they should.

States have slowly begun to adapt new technologies, such as the Internet, to student testing. Just over half the states, for instance, use computers to deliver a portion of the annual state testing programs mandated by the federal No Child Left Behind Act (NCLB).[2] But, for the most part, these states' investments in technology have not led to fundamental changes in our approaches to testing. Mostly, these investments have simply made old approaches to assessment more efficient. Even the most technologically advanced states have done little except replace the conventional paper-based, multiple-choice, fill-in-the-bubble tests with computerized versions of the same.[3] Overall, the types of skills tests measure, and what the test results can tell us, have remained essentially the same.

Technology, however, has the potential to do more than just make our current approach to testing more efficient. A growing number of testing and learning experts argue that technology can dramatically improve assessment—and teaching and learning. Several new research projects are demonstrating how information technology can both deepen and broaden assessment practices in elementary and secondary education, by assessing more comprehensively and by assessing new skills and concepts. All of which can strengthen both national standardized tests like the National Assessment of Educational Progress (NAEP) and classroom-based tests meant to help teachers improve their instruction.

These new technology-enabled assessments offer the potential to understand more than whether a student answered a test question right or wrong. Using multiple forms of media that allow for both visual and graphical representations, we can present complex, multi-step problems for students to solve, and we can collect detailed information about an individual student's approach to problem solving. This information may allow educators to better comprehend how students arrive at their answers and learn what those pathways reveal about students' grasp of underlying concepts, as well as to discover how they can alter their instruction to help move students forward. Most importantly, the new research projects have produced assessments that reflect what cognitive research tells us about how people learn, providing an opportunity to greatly strengthen the quality of instruction in the nation's classrooms. Other

fields, such as military training and medical education, are already using technology-enabled assessment to enhance teaching and learning.

But technology alone cannot transform assessment. Fundamentally changing our approach to testing in our public education system would not be easy. Logistical and funding challenges that often impede efforts to maintain, administer, and update schools' technological infrastructure would have to be overcome. New assessment models must not erode efforts to promote high expectations for all students; nor should they disadvantage low-income schools and students with currently limited access to technology. And new approaches to assessment would have to be aligned with standards, curricula, professional development, and instruction to be successful.

Still, the convergence of powerful new computer technologies and important new developments in cognitive science hold out the prospect of a new generation of student testing that could contribute to significant improvements in teaching and learning in the nation's classrooms.

A Decade of Incremental Progress

Educational researchers and testing experts from around the world have been writing about technology's potential to transform assessment for more than a decade.

The National Academy of Sciences, in its landmark 2001 report, *Knowing What Students Know: The Science and Design of Educational Assessment*, proclaimed it an "opportune time" to fundamentally rethink assessment, citing advances in technology, statistical modeling, and the sciences of thinking and learning.[4] New technology-enabled assessments, supported by research on how students learn, experts argued, would allow us to present complex, multi-step problems and record descriptive data about strategies used and actions taken by students. These data could then be used to adapt instruction by creating a better understanding about students' knowledge, and their conceptual understanding and cognitive development, which would lead not only to better assessment but to significant improvements in instruction and learning.

Similarly, in the late 1990s, Randy Bennett, a scientist at the Educational Testing Service (ETS) who directed the Technology-Based Assessment Project for the National Assessment of Educational Progress (NAEP), predicted that technology would enable educational testing to reinvent itself in three stages. First, technology would increase efficiency by automating existing processes. Secondly, test questions, formats for response, and scoring would become more sophisticated, allowing for the possibility of measuring new skills and measuring currently tested areas more comprehensively. At this stage, Bennett argued, technology would enable a new generation of simulations, and performance assessment would play an essential role. And, thirdly, Bennett envisioned testing merging with instruction, which would allow teachers and students to use feedback from testing to adjust teaching to improve student achievement.[5]

But in the main, the changes that Bennett and the National Academies envisioned have not taken place.

As researchers and cognitive scientists were beginning to recognize information technology's potential, the states were backing away from performance-based assessments, which were designed to mirror more complex, or real-world tasks. In the late 1980s and early 1990s, states began to experiment with using projects, portfolios, exhibitions, and other performance-based activities to measure content mastery.[6] The goal, writes Lorrie Shepard, dean of the University of Colorado's School of Education, was to produce assessments that "more faithfully reflect how learning would be used in non-test situations," assessments that were "guided by an underlying theory of teaching and learning drawn from the cognitive sciences."[7]

But the states' performance assessments were costly and technically inadequate for use in school accountability systems. A 1992 report published by the RAND Corporation on a portfolio assessment program in Vermont found significant problems with the reliability of the program's test scores.[8] It was difficult "to make scores comparable in meaning from year-to-year and from school-to-school," explains Harvard professor and measurement expert Daniel Koretz, who authored the report.[9] States, therefore, began to move away from performance-based assessment systems, back to less-expensive multiple-choice assessments. The demise of large-scale performance-based assessment systems also slowed efforts to link cognitive science with standards-based reform.

The enactment of NCLB in 2002 further complicated attempts to develop new types of testing. NCLB, which mandates that states give annual tests in reading and math in grades 3-8 and once in high school, resulted in a sizeable increase in the number of standardized tests given each year—now more than 45 million—creating a situation in which both test- and policymakers scrambled just to get the tests into the hands of teachers and students.[10] This tremendous increase in test taking, combined with the limited capacity of state departments of education and the nation's testing industry, encouraged state testing officials and testing companies to continue to use the same kinds of tests instead of pursuing innovations in assessment.

And, at key times, NCLB requirements, along with the relative immaturity of new assessment technologies, further slowed the development of new testing models. For instance, while NCLB does not prevent the use of computer-adaptive tests, which adjust the level of question difficulty based on students' answers to previous questions, it does require that tests align with state content standards and that each student be assessed at his or her official grade level.[11] (Lawmakers wanted to ensure that test results would be comparable from student to student and create common standards for all students, regardless of their backgrounds.) By 2002, Idaho and South Dakota were implementing such tests statewide to elementary and secondary students. But these early adaptive tests used by states adjusted to test low-achieving students' performance by presenting items that were below their grade level. Many of the tests were also plagued by technical and content issues.[12] As a result, the U.S. Department of Education would not allow these states to use computer-adaptive tests to meet NCLB requirements.

In 2007, the U.S. Department of Education approved Oregon's request to use a within-grade-level computer-adaptive test for NCLB-mandated state assessments.[13] But, in the meantime, many states signed multi-year contracts and spent millions of dollars investing in traditional, fixed-format tests.[14] The dot-com crash in the early 2000s and resulting state budget shortfalls also dampened enthusiasm for technology-based innovations in student assessment.

Despite this slow progress, assessment experts believe that testing will increasingly be delivered via computer and the Internet—especially as a way to continue to increase the efficiency of testing systems.[15]

Internet-based testing, for instance, eliminates the physical distribution, storage, and collection of test booklets and materials, along with data entry and scanning. Digital delivery and scoring saves time and accelerates the speed with which states can analyze and distribute test results. In 2008, fully 27 states delivered at least one of their state assessment tests via computer.[16] The most prominent, Virginia, administered more than 1.4 million tests online in the spring of 2008.[17]

Computer-adaptive tests offer a different type of efficiency. These tests can produce a more reliable estimate of student achievement using fewer items than required for a traditional test.[18] Since the test quickly adapts to a student's skill level, this form of testing eliminates the need for test items that don't yield information about a student's ability. Answers to easy questions, for example, offer little information to help assess a high-achieving student's specific level. Similarly, overly difficult questions provide little guidance on a low-achieving student's specific level. Whereas a one-size-fits-all test might need to employ 75 test questions to get enough data on students at various levels (for example, 25 questions at low, medium, and high levels), a computer-adaptive test could offer 50 questions instead. Moreover, since these questions are focused on the student's particular achievement level, the test can provide more specific evidence about that student's performance.[19] While computer-adaptive tests are only used for NCLB-mandated assessments in Oregon, they are increasingly used at the district level as practice and benchmark tests.[20]

The efficiencies gained from computer-based testing don't merely reduce the time and money used to administer testing programs. Incorporating automated essay scoring, a technology already in use on standardized tests such as the GMAT, the standardized test used for business-school admissions, enables assessments to test conceptual understanding and writing skills through open-ended, essay responses.[21] In addition, more efficient tests may make it possible to develop more flexible testing programs. Rather than yearly testing, portions of the test could be given throughout the year, offering a more accurate sample of students' progress over time.

For classroom or school-level assessments, results can be made available immediately to teachers, administrators, and district officials. They can also provide a greater connection to instruction, giving educators the chance

to collect information that can be used proactively to inform instruction, rather than only retroactively to gauge success. For example, automated essay scoring allows students to improve drafts with automated feedback. More periodic, flexible, and efficient testing will allow teachers to more easily embed assessment into current instructional processes.

Promising Models

At the same time, a number of promising research projects are beginning to explore the potential of technology to transform testing in more fundamental ways. They suggest that the technology-enabled assessment system that Bennett and others envisioned is indeed possible—a system that's both deeper and broader, able to test knowledge and skills more thoroughly and to test skills and concepts that haven't been measured in the past, and a system that reflects far more fully what we know about how students learn.

Testing Complex Skills

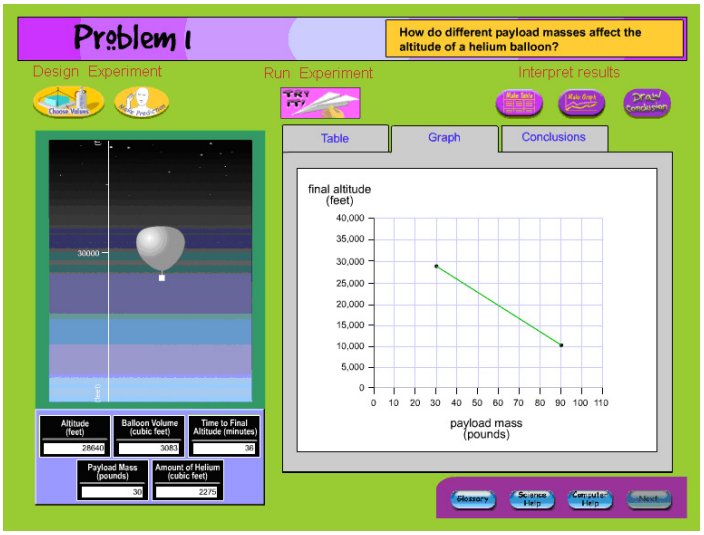
One of the largest efforts to pilot new forms of technology-based assessment is the Problem Solving in Technology-Rich Environments (TRE) project. It was launched in the spring of 2003, when a nationally representative sample of 2,000 students participated in a study to explore how information technology could be incorporated into the country's "gold standard" for assessment—the National Assessment of Educational Progress (NAEP). The goal was to create scenarios "that would feature … the kind of exploration characteristic of real-world problem solving."[22]

TRE tested scientific inquiry skills such as the ability to find information about a given topic, judge what information is relevant, plan and conduct experiments, monitor one's efforts, organize and interpret results, and communicate a coherent interpretation. In one component, eighth-graders used a simulated helium balloon to solve problems of increasing complexity about relationships among buoyancy, mass, and volume. For example, the students were asked to determine the relationship between payload mass and balloon altitude. To solve the problem, students gathered evidence by running simulated experiments using a variety of different payload masses. Once they had enough evidence, they submitted their conclusions using both open-ended and multiple-choice responses.[23]

TRE demonstrates several unique capabilities of technology-enabled assessments. First, technology permits the presentation of more complex, multi-step problems for students to solve. Multiple forms of media, such as the animated helium balloon and instrument panels in TRE, can present information in more useful and compelling ways than text alone. Technology-enabled assessments can present tasks based on complex data sets in ways that even elementary school students can use.[24] In TRE, for example, students see both visual and graphical representations showing what happens to the balloon during each experiment. (See Figure 1.)

Another example of technology-enabled assessment being used in science education is Floaters, a test given to students as part of the World Class Tests optional assessment program in the United Kingdom. The international initiative uses highly visual, engaging questions, enabling young students to be tested on an aspect of scientific method in a way not possible using paper and pencil. Students, for instance, use an interactive simulation to weigh a variety of foods, such as carrots, apples, and bananas, and observe whether the fruit floats in water. Students must then develop a hypothesis about the patterns they observe.[25]

Figure 1. TRE Exercise: The Relationship Between Payload Mass and Balloon Altitude



Source: Randy E. Bennett, Hilary Persky, Andrew R. Weiss, and Frank Jenkins, *Problem Solving in Technology-Rich Environments: A Report from the NAEP Technology-Based Assessment Project* (NCES 2007-466) (Washington, DC: National Center for Education Statistics, U.S. Department of Education, 2007). Retrieved on November 21, 2008 from http://nces.ed.gov/nationsreportcard/pubs/studies/2007466.asp.

Recording More Data

The problems in Floaters and TRE can be dynamic, presenting new information and challenges based on a student's actions. This allows students to take different approaches and even test multiple solutions. And critically, databases can record descriptive data about strategies used and actions taken by students. This provides a greater range of information, allowing instructors to make better judgments about student approaches, challenges, and performance.

In the TRE simulation exercise, for instance, every student action—such as which experiments they ran, which buttons they pushed, and what values they chose and in what order—is logged into a database. (See Figure 2.) Student actions, such as the quality of their experimental design choices, are evaluated using a set of rules and then scored based on statistical frameworks. These algorithms are linked across multiple skills, allowing students to be evaluated based on multiple points of evidence. And since each of the component skills can be traced back to observable student actions, instructors can gather detailed evidence to help determine why a student responded the way they did, helping to identify gaps in skill level, conceptual misunderstandings, or other information that could inform instruction.[26] Instead of just one data point, for example, a right or wrong answer, technology-enabled assessments can produce hundreds of data points about student actions and responses.

Linked to Classroom Instruction

Simulated exercises are particularly useful for assessing students' knowledge of interactions among multiple variables in a complex system, such as in an ecosystem. But, since these models assess both process and content, they require assessments that are closely linked with classroom instruction. This presents a problem for the broad use of these models. TRE, for example, restricted its assessment to scientific problem solving with technology—rather than science content—because NAEP cannot assume that students in the nation's some 14,000 school districts have all covered the same science content. Most of the time in science, however, as University of Maryland researcher Robert Mislevy explains, "it's not 'here's the situation in the world, and you give the answer.' Usually you have some hypotheses, some conjectures, but then you do something, and the world does something back. It's these cycles that really get at the nature of what model-based reasoning under constraints is really about."[27] But with large-scale tests such as NAEP, which Mislevy characterizes as "drop-in-from-the-sky" assessments, "you can't presume anything about what the examinees know about what they're going to be doing."[28]

In contrast, the Calipers project, funded by the National Science Foundation, seeks to develop high-quality, affordable performance assessments that can be used for both large-scale testing and in classrooms to inform instruction. Focused on physical science standards related to forces and motion, along with life sciences standards related to populations and ecosystems, Calipers engages students in problem-solving tasks such as determining the proper angle and speed to rescue an injured skier on an icy mountain. (See Figure 3.) Similar to TRE, Calipers captures descriptive data—describing the approach that a student took to solve the problem (choice of experimental values, formulas chosen), along with multiple-choice and open-ended responses. Calipers hopes to use these descriptive data, along with student reflection and self-assessment activities, to provide information to both students and teachers to guide learning and instruction.[29]

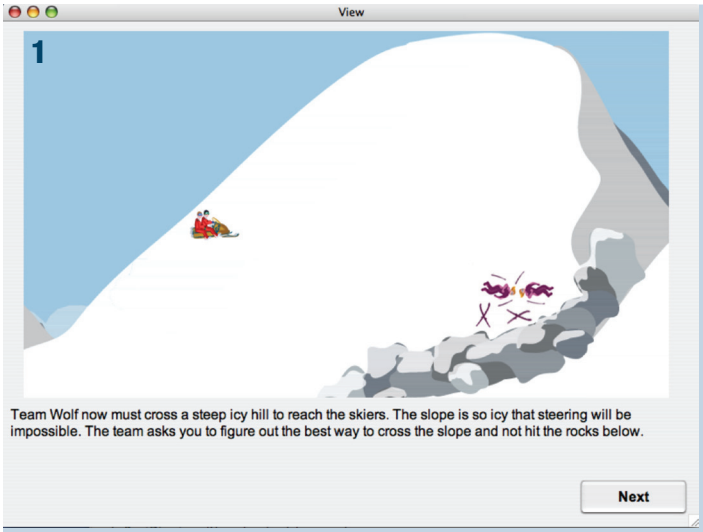
Figure 2. Logging One Eighth-Grader's Actions on a TRE Simulation Exercise (2003)

Time (in seconds)[1]	Action	Action choice
137	Begin problem 1	†
150	Choose values	90
155	Select mass	†
157	Try it	†
180	Make table	†
182	Selected table variables	Payload mass
185	Make graph	†
188	Vertical axis	Altitude
190	Horizontal axis	Helium

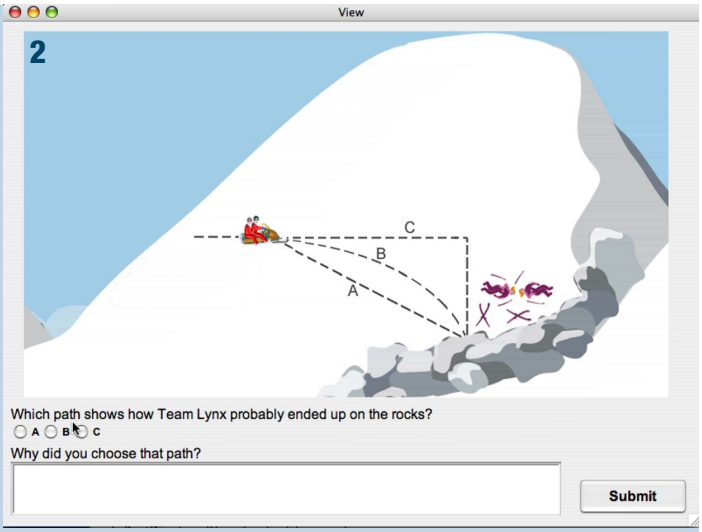
†Not applicable.
[1]These times include 137 seconds spent interacting with introductory material presented prior to problem 1.
Note: TRE=Technology-Rich Environments.

Source: Adapted from Table 2-1. Randy E. Bennett, Hilary Persky, Andrew R. Weiss, and Frank Jenkins, *Problem Solving in Technology-Rich Environments: A Report from the NAEP Technology-Based Assessment Project* (NCES 2007-466) (Washington, DC: National Center for Education Statistics, U.S. Department of Education, 2007). Retrieved on November 21, 2008 from http://nces.ed.gov/nationsreportcard/pubs/studies/2007466.asp.

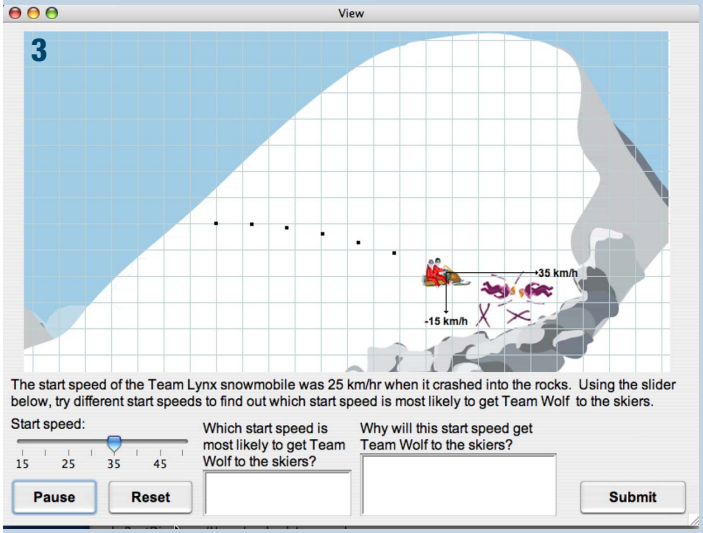
Table 3. CALIPERS Problem: Rescuing Injured Skiers



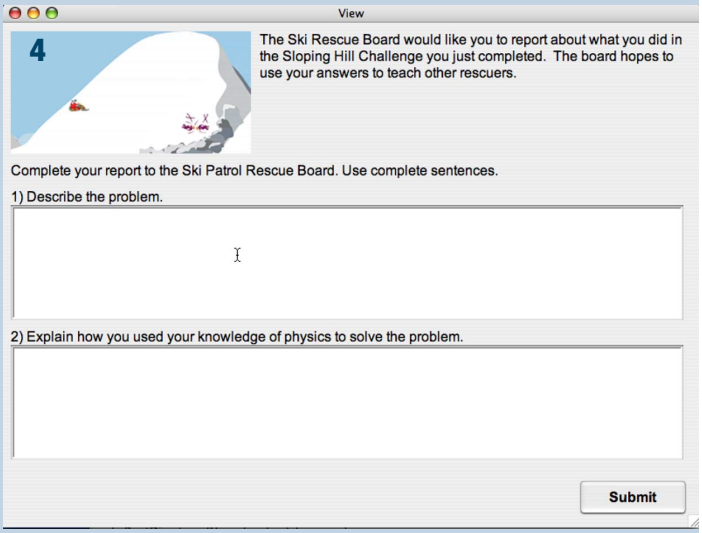
Test-takers are presented with a "real-life" problem that will test their understanding of physics principles.



Test-takers get a chance to choose from multiple options and explain their choice.



Test-takers can manipulate variables to achieve different outcomes, much as they would in the real world.



Test-takers are asked to demonstrate their understanding of the problem and how subject-matter knowledge helped them to solve it.

Source: http://calipers.sri.com/assessments.html.

Right Here in River City

Simulations also provide an opportunity to embed assessment into the learning process. The River City project, led by Harvard education professor Chris Dede, is a multi-user, virtual environment where middle-school students explore a mysterious illness in a turn-of-the-century town. Students learn by becoming scientists in River City's virtual world. With the project focused on inquiry practices, students make observations, "chat" with townspeople, develop hypotheses, and conduct experiments to determine the cause of the epidemic.[30]

Currently, River City uses traditional multiple-choice and teacher-graded assessments, such as a student-written report to the mayor outlining an action plan to eradicate the illness. But, in the future, researchers hope to use these traditional assessments in conjunction with the potential gold mine of descriptive data in the program's database. They are still working to develop algorithms to analyze and make use of the massive volumes of data River City captures about student actions. Jody Clarke, one of the River City researchers, says that the ultimate goal is to present data about what students are doing in the virtual environment in a way that helps teachers organize and

individualize instruction. She also believes that these data can be used to create performance-based summative assessments that are valid, reliable, and cost-effective.[31]

The Cisco Networking Academy, which teaches computer networking skills to 600,000 high school and college students each year in 160 countries around the world, provides another example of assessment that is embedded into learning. Functioning in over 9,000 different schools and 63 developing nations, such as Indonesia, Guinea, Mali, and El Salvador, the academy also demonstrates the potential for technology-enabled assessment at scale and in a wide variety of circumstances and settings.[32] A decade ago, employers complained that students graduating from the academy "do fine on the test, but you put them in front of a busted network and they have no idea what to do."[33] In response, the academy built Packet Tracer, a simulation and assessment engine that enables local instructors to create a variety of simulated computer network environments. With these simulations, students visualize how packets of data move across a network, further their understanding of how a network functions, and test their skills to identify and solve network problems.[34] Driven by a shared desire to assess how students perform in real-life situations, a number of other industries, such as architecture and accounting, are also using computer-based simulation for professional licensure.[35]

Perhaps even more importantly, the Cisco Networking Academy's technology, along with its integration with assessment, curriculum, and instruction, allows the academy to analyze data from tens of thousands of assessments to discover gaps and evaluate enhancements to instruction and curriculum at a program level.[36] Since it is much more defensible to make inferences from assessment data across larger numbers of students, these program-level uses of data are important potential benefits of technology-enabled assessments. Ideally, districts and states could use this type of information to rapidly test interventions across wide ranges of students, leading to the development of a powerful continuous improvement cycle.

Fully immersive simulations, such as those found in medical education and military training, point to further applications of technology. iStan, a life-like, sensor-filled mannequin that can talk, sweat, bleed, vomit, and have a heart attack, is used for medical-training purposes to simulate patient interactions and responses.[37] The U.S. Army has "instrumentalized" many of its war games and other performance exercises, using video and sensors to gather multiple sources of data about what is happening and when. As in the medical school simulations, these extensive data can illustrate multiple interactions among team members. This can lead to productive conversations about what happened, why, and ideas for improvement.[38] These types of assessments and simulated experiences are becoming more prevalent in higher education and the workplace. They engage participants in exercises to problem-solve realistic situations.

This focus on situated assessment, or assessing behavior in realistic situations, is increasingly important at a time when citizens and workers alike need to be able to communicate, collaborate, synthesize, and respond in flexible ways to new and challenging environments. Assessing the ability to approach new situations flexibly is challenging in our current paper-and-pencil environment.[39] "Real-life is not sequestered ….[what is important] is how you respond to feedback, not what you do in a feedback-free world," says John Bransford, University of Washington professor and a leading expert in cognition and learning technology.[40] Bransford is designing assessments that allow students to demonstrate not only what they can recall, but also how they can use their expertise. Technology-enhanced environments and virtual worlds, such as those found in medical training or River City, are necessary for students to practice and gain feedback in real-life, situated environments. In fact, Bransford notes, these efforts are "not possible without technology."

Aligning all the Parts

Education is a complex and decentralized public sector system, funded and governed at multiple levels. As a result, successful changes to assessment will require parallel, and equally challenging, revisions to standards, curriculum, instruction, and teacher training. Without deliberate attention from policymakers and educators in these areas, there is no guarantee that technology will fundamentally change core practices and methods in education, a field that has been notoriously impervious to change. Stanford University education historian Larry Cuban cautions that the "persistent dream of technology driving school and classroom changes has continually foundered in transforming teaching practices."[41] Just adding technology and hoping for educational transformation, without considering the content and practice of instruction, will do no more than automate existing processes, Cuban argues.

Standards and Cognitive Models

The cognitive research presented in the National Academy of Sciences 2001 report *Knowing What Students Know* stresses the importance of aligning assessments with curriculum and instruction and the need to base testing on a model of cognition and learning.[42] Yet, most state standards, as currently developed, focus on discrete sets of disconnected facts.[43] They do not provide a clear sense of where students are relative to desired goals, nor do they provide a complete road map for students and teachers to navigate.[44] In other words, our assessments do not align with what we know about how students learn and do not tell us enough about how to help students do better.

The disconnect is most evident in science education. Increasing global competition, a changing economy, and years of mediocre test results on international comparisons have sparked broad agreement among policymakers and educators that U.S. students must improve in the science, technology, engineering, and math (STEM) subject areas.[45] As such, many recognize that a different approach to teaching, learning, *and* assessment is needed. The National Academy of Sciences, in its 2007 report *Taking Science to School: Learning and Teaching Science in Grades K–8*, calls for "a redefinition of and a new framework for what it means to be proficient in science."[46] Their framework for science education, based on research on how students learn science, states that "content and process are inextricably linked in science." Scientific practices, such as inquiry, cannot be taught in isolation from the underlying concepts. But our tests, whether paper-based or online, focus almost exclusively on factual knowledge.

Also, while multi-user environments, simulations, and other technological domains offer many capabilities and opportunities, these tools are only as good as the cognitive models on which they are based. Mislevy, of the University of Maryland, cautions that "the evidentiary foundation … must be laid if we are to make sense of complex assessment data."[47] We can't use the data that these tools generate to inform assessment and instruction unless we have a greater understanding of how students learn within a domain. In a forthcoming article with fellow researcher Drew Gitomer, ETS' Bennett explains, "In principle, having a modern cognitive-scientific basis should help us build better assessments in the same way as having an understanding of physics helps engineers build better bridges."[48]

In fact, technology-enabled assessments expose the flaws in our current development of educational standards.[49] Most standards, for instance, are written as if we've asked teachers to ensure that their students can drive to a specific destination … let's say, Albuquerque. Our current assessments can tell us if a student has arrived, but don't tell us whether the students who haven't arrived are on their way, made a wrong turn, or have a flat tire. Technology-enabled assessments could in principle be like a GPS system, with the capability to frequently monitor and assess progress along the way. But a GPS is useless without the software that relates physical latitude and longitudinal coordinates back to a detailed map complete with roads, possible detours, and routes to Albuquerque. Similarly, to be transformative and to enhance teaching and learning, technology-enabled assessments will need to be dependent on a detailed understanding of how learning progresses in math, science, and various other disciplines. So far, however, our technological capabilities surpass our knowledge in these areas.[50]

For example, while there is much potential in the use of computer-based, multi-player gaming for both learning and assessment, we don't fully understand how gaming activities connect with the learning outcomes we are trying to teach and assess. Sigmund Tobias, a research scientist at Teachers College, Columbia University and J.D. Fletcher, a senior researcher at the Institute for Defense Analyses, in their review of research on gaming and learning, argue that even though a game may have similarities to what is being taught or assessed for real-life use, the important learning outcomes don't necessarily transfer. It's essential to analyze the actual cognitive tasks involved in the game and map them to the goals, they write.[51]

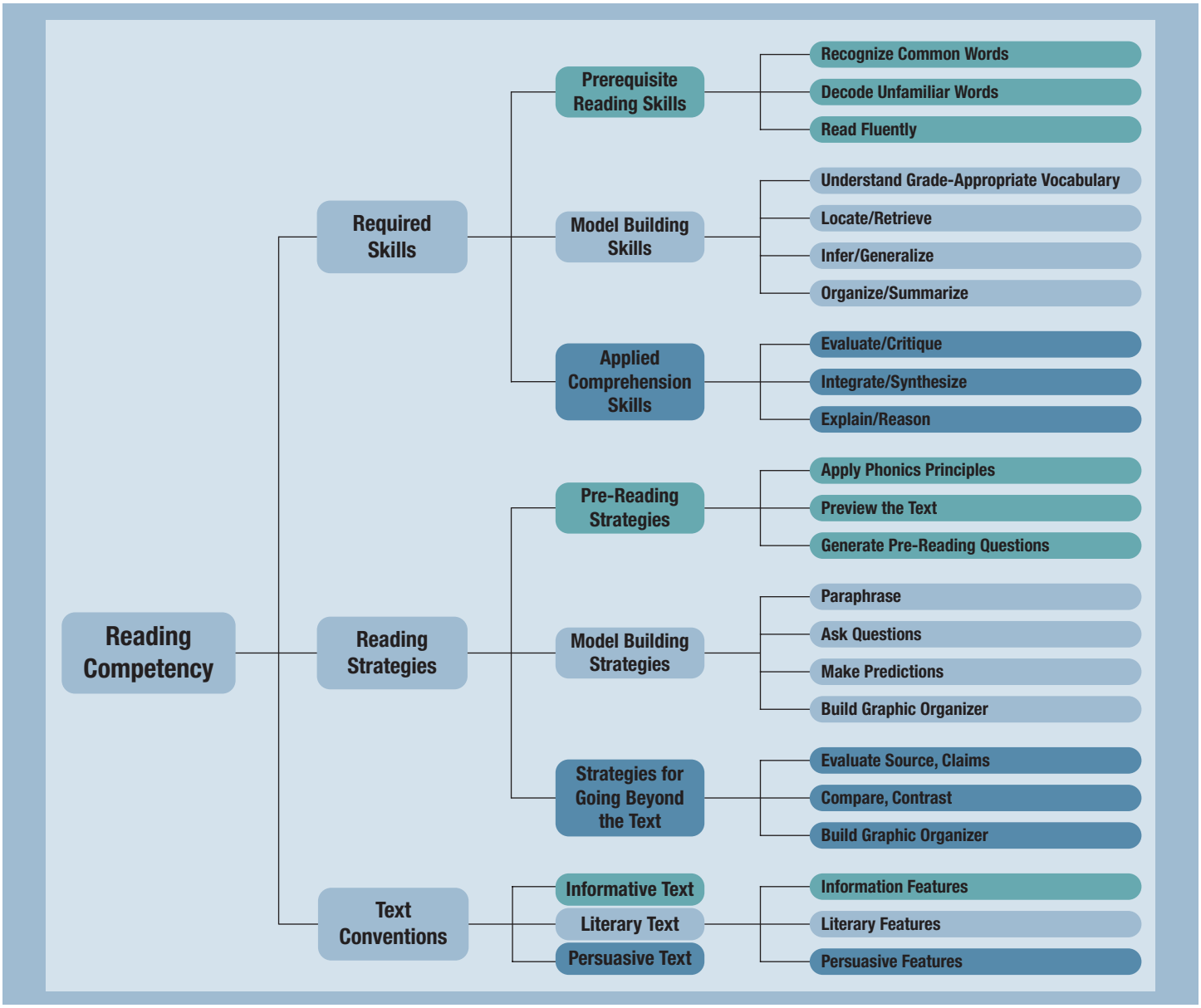
The Educational Testing Service's Cognitively Based Assessment of, for and as Learning (CBAL) research project provides an example of where both technology and the map come together. While the project is dependent on technology—it uses automated essay scoring and the research takes place in Portland, Maine, in schools with one-to-one laptop programs—the extensive research and development of a cognitive model for how students read and develop reading skills is also essential.

The cognitive model forms the bridge between two different uses of testing. Summative assessment describes what has been learned over time and is used to judge the performance of students or schools,

while formative assessment is meant to guide teaching and learning as part of the instructional process.[52] Projects built on cognitive models, such as CBAL and Calipers, attempt to build both summative and formative components, held together by a common conception of how students learn a particular subject.

For example, in CBAL, the model for reading competency includes three broad categories of required skills: pre-requisite reading skills, model building skills, and applied comprehension skills.[53] (See Figure 4.) Each of these categories is fully developed in the model and assessed during the periodic tests given over the course of a school year. Questions test applied literacy, with tasks such as evaluating text content for bias and evidence to support claims using a wide variety of sources, such as newspaper articles, encyclopedia entries, and diagrams. The cognitive model underlying CBAL ensures that the

Figure 4. Cognitive Model for Reading Assessment



Source: K. M. Sheehan and T. O'Reilly, "The Case for Scenario-Based Assessments of Reading Competency" (paper presented at the Assessing Reading in The 21st Century Conference: Aligning and Applying Advances in the Reading and Measurement Sciences, Philadelphia, PA., April 2008). Reprinted by permission of Educational Testing Service, the copyright owner. No endorsement of this publication by Educational Testing Service should be inferred.

project's summative assessments, meant to be used for accountability purposes, accurately align with and assess all of the various dimensions of reading. Still in its early stages, this conceptual model has also allowed CBAL researchers to begin differentiating students and instructional responses to those students based on their performance on the CBAL assessments.[54]

Similarly, research in the United Kingdom is testing how technology-enabled assessments, combined with advanced statistical and cognitive models, allow teachers to identify groups of readers with different patterns of performance even though the students' raw test scores may be similar. Teachers can then tailor instruction to four types of readers—reluctant readers, developing readers, reasoning readers, and involved readers.[55]

But without a sound evidentiary model and conceptual underpinning, technology-enabled assessment tools are just more efficient, faster, and accessible versions of the same old tests. For example, although many Internet-based benchmark tests are marketed as formative assessment products, most do not provide the specific information needed to improve instruction. Increasingly popular at the district level, these online test systems come equipped with banks of test questions for classroom teachers to use. However, rather than drawing from a common cognitive model, they are essentially just online variations of the types of questions found on state assessments. They provide practice on these question types and predict what a student will score on a state assessment test, but they don't tell a teacher *why* a student scored a certain way or how to modify instruction.

Effective Instruction

The most useful classroom-level applications of technology-enabled assessments, such as CBAL, River City, and others, provide descriptive data that give teachers better information about how students are progressing and why they are performing at their current levels. They also provide insight into what changes to instruction might be the most effective.

These formative uses of technology-enabled assessments are both promising and challenging. Formative assessment, when implemented effectively, is one of the few reforms to show significant effects on student gains—gains that research shows to be especially pronounced for the most struggling students.[56]

But this classroom-level use is highly dependent on effective teaching. Teachers must have deep content expertise and specific skills to understand and use the insights these sophisticated assessments produce. The teachers must also be open to an entirely new approach to instruction. David Niemi, formerly with UCLA's National Center for Research on Evaluation, Standards, and Student Testing, writes, "Teachers must achieve not only deep understanding of the subjects they teach but also broad knowledge about how that understanding typically develops over time and how it may be assessed."[57]

Andrew Boyle, a leading researcher on technology-enabled assessment (known as "e-assessment" in the United Kingdom), cautions that "at the present time, sophisticated tasks for e-assessment may be characterized as expensive and slow to develop, and not easily written by a non-specialist teacher."[58] Technology has enabled mass customization in a number of areas, but the challenge of creating high-quality assessments and simultaneously making them adaptable by teachers for easy classroom use is considerable.

Infrastructure and Equity

In addition to the teacher-quality challenge, most school technology infrastructures are still insufficient for advanced, large-scale, Internet-based testing. In *Computer-Based Assessment: Can It Deliver on Its Promise*, a 2001 policy brief published by WestEd, a nonprofit research agency, the authors concluded, "It seems only a matter of time before computer-based assessment defines the test administration process."[59] Yet, in a November 2008 feature on computer-based assessment in *Education Week*, the writer noted that "progress toward that goal has been slow, expensive, and fraught with logistical challenges."[60] Inadequate computer hardware, bandwidth constraints, and limited capacity to maintain, update, and administer schools' technological infrastructure have proven to be serious impediments to deploying online testing, especially for more secure, high-stakes state NCLB testing.

The serious challenges encountered in 2008 by NAEP in its initial testing of computer-based questions for the 2009 Science Assessment provide an instructive case study.

Initially, NAEP administrators relied on schools' computers and technology infrastructure to administer the student exams. However, due to a wide variety of problems with the schools' hardware, software configurations, and Internet access, technical problems impeded half the students in the first few weeks of the trial. Finally, NAEP resorted to bringing its own laptops into each school and equipped every student with a secure portable flash drive to store data. This experience, along with the tremendous logistical burdens and other costs associated with transporting and setting up laptops in each of the schools, led NAEP to significantly scale back its plans for technology-enabled testing in 2009. Interactive, technology-enabled questions will only be given to a small subset of schools participating in the 2009 NAEP.[61]

Moreover, concerns that technology-enabled assessment will disadvantage primarily low-income students that may have limited access to computers and other forms of technology must also be overcome. For example, if an assessment that is developed to assess math skills inadvertently tests computer skills, then the results will be biased against students with less computer access. Results from earlier NAEP technology-enabled assessment projects, Math Online and Writing Online, found that students with greater exposure to computers outside of school scored higher.[62] Other studies show no difference between paper and computer-based tests, indicating that more research is needed to understand how to avoid bias against students without computer familiarity.[63] At the same time, technology can help address the needs of specific populations, such as English-language learners or students with disabilities. (See sidebar, "The Benefits of Universal Design," page 12.)

Changing Course

Overcoming these barriers will be challenging. But so is the goal of helping all students reach challenging standards for learning and be prepared for future success. To reach these goals educators must be able to adapt instruction to account for a multitude of variables. This adaptive instruction is not possible without a deeper understanding of and better data about how students learn.[64]

Now with the impending reauthorization of NCLB, there's an opportunity to begin to chart a different course for the future of educational assessment, one that would also enhance teaching and learning, which is essential to meeting NCLB's goals. In fact, by requiring state officials to rapidly expand testing in American classrooms, NCLB helped bring the inadequacy of our current testing practices to the forefront and created a rare consensus among proponents, critics, teachers and policymakers—none are satisfied with the state of testing today.

With technology changing at a rapid pace, we have many of the tools to create vastly improved assessment systems and practices. Dozens of research projects, both here in the United States and in countries around the world, are beginning to demonstrate how technology, used in concert with what cognitive-scientific research tells us about how people learn, can improve formative assessment and enhance teaching and learning. And, the experiences from a variety of industries, ranging from medicine to accounting, shows that simulations and more advanced, technology-enabled assessments can also be used for large-scale and consequential testing.

And despite the infrastructure challenges, NAEP, which is perhaps our nation's most important large-scale test, is moving quickly to incorporate technology into its main assessment program. With a main key goal to "go beyond what can be measured using paper-and-pencil," the 2009 NAEP Science Assessment will use technology-based assessment tasks for a subsample of the students taking the test. In 2011 NAEP plans to go even further—that year's eighth-grade writing assessment will be delivered entirely by computer.[65]

Yet, because changes in assessment impact our entire educational system and infrastructure, from state agencies to test-makers to federal officials to classroom teachers, we won't see the real benefits from technology-enabled assessments—improved teaching and learning—without careful attention from policymakers and deliberate strategies to create change. It will take time to further research, build, and implement on a wide-scale the types of technology-enabled assessments described in this report. But in the meantime there are steps that policymakers, educators, and a variety of stakeholders can take immediately to ensure that progress moves much more quickly in the next decade.

To begin with, if we want to see dramatically better assessments, we can't continue to invest solely in the

The Benefits of Universal Design

On television sets all over the country, closed-captioning decoder chips allow exercisers on treadmills, fans at noisy sports bars, students learning a second language, and hearing-impaired persons to follow the dialogue on their favorite television shows. But decoder chips weren't always built into every television. Prior to built-in chips, hearing-impaired persons had to seek special accommodations, such as expensive set-top decoder boxes, to access closed captioning. Integrating decoder chips into every television not only offered much greater access to the hearing-impaired, but was highly cost-effective and provided valuable benefits to a wide range of television viewers.

Universal design, the concept behind this innovation, allows designers to address the divergent needs of special populations and at the same time, increase usability for everyone.[a] Experience from a wide variety of fields, from architecture to civil engineering to consumer products, shows that the application of universal design does not only provide wide benefits, but can save money by preventing costly retrofits and special accommodations.

Technology allows developers to apply these same universal design concepts to educational assessment. Digital materials are inherently flexible, making it feasible to customize materials and methods to each individual.[b] Rather than provide special accommodations, such as a separately printed test booklet with enlarged print for students with reduced vision, technology-enabled testing can embed a variety of accommodations into the same program. In an *Education Week* commentary, Andrew Zucker, author of *Transforming Schools With Technology* and a senior research scientist at the Concord Consortium, writes that "Computers enlarge typographical fonts, translate to and from English, convert text to speech, correct mistakes, and help teachers individualize instruction."[c] Applying universal design to assessment offers the potential to allow all students to benefit from a testing environment that adapts to meet individual needs.[d]

Other, more targeted accommodations are also possible. Researchers are investigating the use of graphical interfaces to make tests more accessible and valid for specific populations. It is difficult, for instance, to test English-language learners, students with low English proficiency, in subjects such as science without mistakenly testing their language skills instead. Rebecca Kopriva, a University of Wisconsin researcher, is currently testing computer-based science assessments that use animations and non-verbal communications to test concepts such as what happens to water molecules when they are heated.[e]

But, to be effective, these accommodations must go beyond just assessment. Accommodations built into technology-enabled assessments must also be available for students throughout the entirety of their educational experiences. In fact, the National Center for the Improvement of Educational Assessment notes that "If the first time a student sees a certain type of accommodation—especially if it is somewhat novel—is on the large-scale assessment, that accommodation will likely hinder instead of improve access."[f] As in every other aspect of assessment design, usability enhancements to assessment must also align with similar enhancements to curriculum and instruction.

[a] David H. Rose, Anne Meyer, Nicole Strangman, and Gabrielle Rappolt, *Teaching Every Student in the Digital Age* (Alexandria, VA: Association for Supervision and Curriculum Development, 2002). Retrieved December 10, 2008 from http://www.ascd.org/publications/books/101042.aspx.

[b] National Center on Accessing the General Curriculum, "Virtual Reality and Computer Simulations and the Implications for UDL Implementation: Curriculum Enhancements Report." Retrieved December 10, 2008, from http://www.k8accesscenter.org/training_resources/udl/documents/VirtualRealityUDL_000.pdf.

[c] Andrew Zucker, "Commentary: Smart Thinking About Educational Technology," *Education Week*, April 2, 2008.

[d] G. Madaus, M. Russell, and J. Higgins, *The Paradoxes of High-Stakes Testing: How They Affect Students, Their Parents, Teachers, Principals, Schools, and Society*.

[e] "Computer Simulations: Four Approaches to Interactive Student Assessment" (presentation at the Council of Chief State School Officers National Conference on Student Assessment, June 17, 2008).

[f] Marianne Perie and Scott Marion, "Computer Based Testing in Utah: A Summary of Key Issues and Stakeholder Input," an unpublished report from the Center for Assessment, February 29, 2008.

current practice. In addition to today's current federal dollars for assessment, federal policymakers should create a second, smaller pool of funding to support the research and development of the next generation of assessment technology and practice. This money should be focused not just on isolated research projects, but on applied applications that involve competitive awards and partnerships among researchers, psychometricians, testing companies, state officials, and educators. Rather than just award contracts, funds should support both early developmental work and provide large incentives for actual district or state-level implementation.

Also, because the cognitive and data models underlying technology-enabled assessments will be used to gauge student progress and guide instruction as much as possible and much more than today, they must be open for public review and improvement, ensuring that evaluators can test and enhance these models along the way. And, within the broad parameters of federal policy and coupled with rigorous evaluation, schools and educators participating in these innovative initiatives should have the freedom to use, as well as incentives for the use of, new assessments and assessment practices—both summative and formative. Given the importance of

science, the deficiency of current science assessment practices, and the number of promising research projects in this field, science education is a logical place to start.

We should plot a different course—one that maintains accountability goals but creates space for significant innovation and prioritizes the use of technology-enabled assessments not just for automation, but for substantive improvements in student achievement.

Endnotes

[1] Retrieved on September 25, 2008, from http://www-03.ibm.com/ibm/history/exhibits/specialprod1/specialprod1_9.html

[2] C.V. Bausell, "Tracking U.S. Trends," *Education Week*, March 27, 2008. For more on states' efforts to implement online testing, please see Katie Ash, "States Slow to Embrace Online Testing," *Education Week*, November 19, 2008.

[3] G. Madaus, M. Russell, and J. Higgins, *The Paradoxes of High-Stakes Testing: How They Affect Students, Their Parents, Teachers, Principals, Schools, and Society* (Charlotte, NC: Information Age Publishing. Forthcoming).

[4] James W. Pellegrino, Naomi Chudowsky, and Robert Glaser, eds., *Knowing What Students Know: The Science and Design of Educational Assessment*, Committee on the Foundations of Assessment, Board on Testing and Assessment, Center for Education, National Research Council (Washington, DC: The National Academies Press, 2001).

[5] Randy Bennett, *Reinventing Assessment: Speculations on the Future of Large-Scale Educational Testing* (Princeton, NJ: Educational Testing Service, 1998).

[6] Daniel Koretz, *Measuring Up: What Educational Testing Really Tells Us* (Cambridge, MA: Harvard University Press, 2008).

[7] Lorrie Shepard, "A Brief History of Accountability Testing: 1965-2007," in The Future of Test-Based Educational Accountability, eds. Katherine E. Ryan and Lorrie A. Shepard (New York: Routledge, 2008).

[8] Daniel Koretz, Daniel F. McCaffrey, Stephen P. Klein, Robert M. Bell, Brian M. Stecher, "The Reliability of Scores from the 1992 Vermont Portfolio Assessment Program" (Santa Monica, CA: RAND Corporation, 1992). Retrieved on September 22 from http://www.rand.org/pubs/drafts/DRU159/

[9] Daniel Koretz, *Measuring Up: What Educational Testing Really Tells Us*.

[10] See Thomas Toch, *Margins of Error: The Education Testing Industry in the No Child Left Behind Era* (Washington, DC: Education Sector, 2006) for more discussion of the effects of the rapid increase in standardized testing.

[11] Rhea R. Borja, "South Dakota Drops Online 'Adaptive' Testing," *Education Week*, January 29, 2003.

[12] Ibid.

[13] As of December 2008, Oregon remains the only state approved to use computer adaptive testing for NCLB assessments. In a letter to the Utah State Superintendent of Public Instruction, dated November 17, 2008, Assistant Secretary of Education Kerri L. Briggs denied a waiver for a pilot computer adaptive program in Utah. "It is vital that all students are validly and reliably tested on assessments aligned to a state's challenging academic content standards in order to obtain an accurate measure of student performance." The state "has not demonstrated that the pilot assessments comply with ESEA assessment requirements, including those for technical quality."

[14] G. Madaus, M. Russell, and J. Higgins, *The Paradoxes of High-Stakes Testing: How They Affect Students, Their Parents, Teachers, Principals, Schools, and Society*.

[15] Katie Ash, "States Slow to Embrace Online Testing," *Education Week*, November 19, 2008.

[16] C.V. Bausell, "Tracking U.S. Trends." *Education Week*, March 27, 2008.

[17] "eTesting at the State Level—Lessons Learned" (presentation at the Council of Chief State School Officers conference, June 16, 2008).

[18] G. Madaus, M. Russell, and J. Higgins, *The Paradoxes of High-Stakes Testing: How They Affect Students, Their Parents, Teachers, Principals, Schools, and Society*.

[19] Computer-adaptive tests are most efficient when used for selection or as a gauge of overall ability, for instance, on the GRE. They do not provide as many scoring efficiencies in a standards-based environment if educators are attempting to understand how well a student did in each particular aspect of a subject, for instance, math, because then the test needs to ask questions in each specific aspect. Researchers continue to caution that there are still limitations to computer adaptive testing, such as concerns about whether scores are well-estimated for all examinees and the need for large item pools to ensure test security.

[20] For more information on computer-adaptive testing, please see Katie Ash, "Adjusting to Test Takers," *Education Week,* November 19, 2008.

[21] On the GMAT, a human reader also scores each essay, and additional reviews are conducted if the scores between the computer and human raters disagree significantly. See Greg Miller, "Computers as Writing Instructors," *SCIENCE*, January 2, 2009. Findings across a number of studies consistently show comparability of human and computer scoring at levels approximating the agreement between two human scorers. See Edys Quellmalz and James Pellegrino, "Technology in Testing," *SCIENCE*, January 2, 2009.

[22] Randy E. Bennett, Hilary Persky, Andrew R. Weiss, and Frank Jenkins, *Problem Solving in Technology-Rich Environments: A Report from the NAEP Technology-Based Assessment Project* (NCES 2007-466) (Washington, DC: National Center for Education Statistics, U.S. Department of Education, 2007). Retrieved on November 21, 2008 from http://nces.ed.gov/nationsreportcard/pubs/studies/2007466.asp.

[23] Ibid.

[24] Jim Ridgway and Sean McCusker, "Using Computers to Assess New Educational Goals," *Assessment in Education* 10, no. 3 (November 2003).

[25] Andrew Boyle, "Sophisticated Tasks in E-Assessment: What Are They and What Are Their Benefits?" (paper presented at 2005 International Computer Assisted Assessment Conference, Loughborough University, United Kingdom). Retrieved on July 17, 2008 from http://www.caaconference.com/pastConferences/2005/proceedings/BoyleA2.pdf

[26] Randy E. Bennett, Hilary Persky, Andrew R. Weiss, and Frank Jenkins, *Problem Solving in Technology-Rich Environments: A Report from the NAEP Technology-Based Assessment Project*.

[27] Robert Mislevy, University of Maryland, personal communication, February 1, 2008.

[28] Ibid.

[29] Quellmalz, Edys S., et al. "Exploring the Role of Technology-Based Simulations in Science Assessment: The Calipers Project" in *Assessing Science Learning: Perspectives From Research and Practice*, eds. Janet Coffey, Rowena Douglas, and Carole Stearns (Arlington, VA: National Science Teachers Association Press, 2008); "Computer Simulations: Four Approaches to Interactive Student Assessments" (presentation at the Council of Chief State School Officers conference, June 17, 2008).

[30] Christopher Dede, "Transforming Education for the 21st Century: New Pedagogies that Help All Students Attain Sophisticated Learning Outcomes," Commissioned by the NCSU Friday Institute, February, 2007. Retrieved from http://www.gse.harvard.edu/~dedech/Dede_21stC-skills_semi-final.pdf

[31] Jody Clarke, Harvard Graduate School of Education, personal communication, March 14, 2008.

[32] Statistics provided by Cisco Networking Academy, retrieved from http://www.cisco.com/web/learning/netacad/academy/index.html on December 5, 2008.

[33] Robert Mislevy, University of Maryland, personal communication, February 1, 2008.

[34] John Behrens, director, Networking Academy Learning Systems Development, Cisco Systems, Inc., personal communication, April 2, 2008.

[35] See, for example, the computer-based simulations used in the United States Medical Licensing Examination Step III test. Other examples include architecture (NCARB) and accountancy (AICPA); Randy Bennett, Educational Testing Service, personal communication, November 21, 2008.

[36] John Behrens, personal communication.

[37] Personal demonstration, Games, Learning, and Society Conference, Madison, Wis., July 11, 2008; "Remote-controlled 'man' Called IStan Trains Healthcare Professionals," *Medical News Today*, July 28, 2008. Retrieved on August 20, 2008, http://www.medicalnewstoday.com/articles/116286.php

[38] Dexter Fletcher, Institute for Defense Analyses, personal communication, July 09, 2008.

[39] John D. Bransford, Ann L. Brown, and Rodney R. Cocking, eds., *How People Learn: Brain, Mind, Experience, and School*, Committee on Developments in the Science of Learning, National Research Council (Washington, DC: The National Academies Press, 1999).

[40] John Bransford, personal communication, March 19, 2008.

[41] Larry Cuban, "Techno-Reformers and Classroom Teachers," *Education Week*, October 9, 1996.

[42] James W. Pellegrino, Naomi Chudowsky, and Robert Glaser, eds., *Knowing What Students Know: The Science and Design of Educational Assessment*, Committee on the Foundations of Assessment, Board on Testing and Assessment, Center for Education, National Research Council (Washington, DC: The National Academies Press, 2001).

[43] Richard A. Duschl, Heidi A. Schweingruber, and Andrew W. Shouse, eds., *Taking Science to School: Learning and Teaching Science in Grades K–8*, Committee on Science Learning, Kindergarten through Eighth Grade, (Washington, DC: National Academy of Sciences, 2007).

[44] Margaret Heritage, "Learning Progressions: Supporting Instruction and Formative Assessment," (paper prepared for the Council of Chief State School Officers, 2008).

[45] See "Technology Counts: The Push to Improve Science, Technology, Engineering, and Mathematics," *Education Week*, March 27, 2008, for more background on calls for improved science education,

[46] Richard A. Duschl, Heidi A. Schweingruber, and Andrew W. Shouse, eds., *Taking Science to School: Learning and Teaching Science in Grades K–8*.

[47] Robert J. Mislevy, "Leverage Points for Improving Educational Assessment," U.S. Department of Education Office of Educational Research and Improvement Award #R305B60002, National Center for Research on Evaluation, Standards, and Student Testing (CRESST) Center for the Study of Evaluation (CSE) Graduate School of education and Information Studies, University of California, Los Angeles, 2000.

[48] Randy E. Bennett and Drew H. Gitomer, "Transforming K–12 Assessment," in *Assessment Issues of the 21st Century*, eds. C. Wyatt-Smith and J. Cumming (New York: Springer Publishing Company, Forthcoming).

[49] Chris Brown, senior vice president, Pearson Education, personal communication, March 12, 2008.

[50] Derek Briggs, professor, University of Colorado, personal communication, April 18, 2008.

[51] Sigmund Tobias and J.D. Fletcher, "What Research Has to Say About Designing Computer Games for Learning," *Educational Technology* 47 (2007).

[52] Definitions taken from "Building Balanced Assessment Systems to Guide Educational Improvement" Doris Redfield, Ed Roeber, and Rick Stiggins, background paper for the National Conference on Student Assessment, June 15, 2008.

[53] Tenaha O'Reilly and Kathleen M. Sheehan, *Cognitively Based Assessment of, for, and as Learning: A 21st Century Approach for Assessing Reading Competency* (Princeton, NJ: Educational Testing Service, 2008). Retrieved November 21, 2008, from http://www.cogsci.rpi.edu/csjarchive/proceedings/2008/pdfs/p1613.pdf

[54] Drew Gitomer, researcher, Educational Testing Service, personal communication, March 11, 2008; also see ibid above.

[55] Chris Whetton and Marian Sainsbury, National Foundation for Educational Research, UK, "E-Assessment for Improving Learning" (paper presented at 33rd International Association for Educational Assessment, Baku, Azerbaijan, September 2007).

[56] Dylan Wiliam, "Changing Classroom Practice,", *Educational Leadership* 65, no. 4 (Dec./Jan. 2008).

[57] David H. Niemi, "Cognitive Science, Expert-Novice Research, and Performance Assessment," *Theory into Practice* 36, no. 4, New Directions in Student Assessment, (Autumn 1997).

[58] Andrew Boyle, "Sophisticated Tasks in E-Assessment: What Are They and What Are Their Benefits?"

[59] Stanley Rabinowitz and Tamara Brandt, Computer-based Assessment: Can it Deliver on its Promise? (San Francisco, CA: WestEd, 2001).

[60] Katie Ash, "States Slow to Embrace Online Testing."; see also Andrew Trotter, "Online Testing Demands Careful Planning," *Education Week Digital Directions*, June 20, 2007, for further discussions of challenges faced by states implementing online testing, including Virginia.

[61] "At Last! Computer Based Testing for Main NAEP," (presentation at the Council of Chief State School Officers National Conference on Student Assessment, June 17, 2008).

[62] Randy E. Bennett, J. Braswell, A. Oranje, B. Sandene, B. Kaplan, and F. Yan, "Does It Matter If I Take My Mathematics Test on Computer? A Second Empirical Study of Mode Effects in NAEP," *Journal of Technology, Learning, and Assessment* 6, no. 9 (2008). Retrieved November 10, 2008 from www.jtla.org. Also, N. Horkay, R. E. Bennett, N. Allen, B. Kaplan, and F. Yan, "Does It matter If I Take My Writing Test on Computer? An Empirical Study of Mode Effects in NAEP," *Journal of Technology, Learning and Assessment* 5, no. 2 (2006). Retrieved November 21, 2008 from http://escholarship.bc.edu/jtla/vol5/2/

[63] See Vol. 6 of *The Journal of Technology, Learning and Assessment* (JTLA), www.jtla.org, for additional studies and more discussion of this issue.

[64] For more discussion of assessment's role in adapting instruction, see Thomas B. Corcoran and Frederic A. (Fritz) Mosher, *The Role of Assessment in Improving Instruction* (Philadelphia, PA: Consortium for Policy Research in Education (CPRE) Center on Continuous Instructional Improvement (CCII), October, 2007).

[65] "At Last! Computer Based Testing for Main NAEP," (presentation at the Council of Chief State School Officers conference, June 17, 2008).

Beyond the Bubble: Technology and the Future of Student Assessment