knowledge *brief*

COMPUTER-Based ASSESSMENT

Can it *deliver* on its *promise?*

It's April, and a class full of students is about to tackle the rigorous, high-stakes statewide assessment. Last year they had all sharpened their number 2 pencils, sat down at their desks, opened thick test booklets, and begun filling in the blanks on paper answer booklets. This year, these students are sitting in the school's computer lab, taking their exam on computers connected to the Internet. Two of them, a visually impaired boy and a girl who has severe dyslexia, wear headphones, working at computers that will read the test aloud. Other students in the school are scheduled to take the test at other times in the coming weeks. Instead of answering page after page of multiple-choice questions, all these students will write essays, graph math solutions, and use various computer-based tools to solve real-world problems.

Illustration by James Yang

The computers immediately score students' responses, including their answers to essay questions. Based on how a student answers particular questions, the computer branches to harder or easier questions, an adaptive process that yields more detailed information about what a student does or doesn't know. Because the tests are adaptive, students can also answer fewer questions than needed on a traditional paper assessment, yet the test yields more specific results. When each student is done, his or her test scores are automatically emailed to the teacher and principal. Exam data, sent via the Internet, are stored in the main computer at the state department of education, readily available to schools as needed.

Written by
Stanley Rabinowitz and Tamara Brandt

WestEd
Improving education through research, development, and service

To some, this vision of a more efficient and informative assessment process sounds like a fantasy. But many states don't think so, and they are actively exploring the conversion of their statewide assessment systems from pencil-and-paper exams to computer-based assessments. In turn, test publishers are working feverishly to stake their claims on that territory. Bold statements abound about the promise of computer-based testing.

Can technology-assisted assessment live up to these promises?

Throughout the relatively brief history of high-stakes assessment, many innovations have come along that promised to revolutionize the assessment experience. Some delivered. For example, innovations in the scoring of essays (e.g., scoring rubrics, large scoring centers) have truly changed the face of assessment, leading to the large majority of programs (state, local, National Assessment of Educational Progress) incorporating direct writing or constructed-response questions. Other innovations, such as performance-based methods (e.g., portfolios, projects), have proven harder to sustain and have been relegated to low-stakes smaller programs or local assessments.

Now, with computer-based assessment, comes the possibility of radically improving both how assessments are implemented and the quality of the information they can deliver. But as many states consider whether to embrace the new technologies — and as some already have — serious concerns remain about the fairness of the new systems and the readiness of states (and their districts and schools) to support them. This *Knowledge Brief* first describes the potential advantages of a fully implemented computer-based assessment system. It ends more cautiously, laying out a series of questions states must address as they consider the next generation of high-stakes assessment.

Technology is no stranger to assessment. In the middle of the last century, the rise of multiple-choice methodology for large-scale assessment was fueled heavily by the development of high-speed scanners. More recently, computer-adaptive models, such as those described in the opening vignette, where students are presented with questions tailored to their ability levels, have promised to make assessment more efficient and able to target the needs of individual students. But past advances pale compared to those of the last decade, which has seen a rapid increase in both the use and potential of technology to support assessment. On the hardware side, advances in the speed, capacity, and availability of computers allow applications that could only be imagined less than a generation ago. On the software side, developments in database structures, simulation technologies, and artificial intelligence models promise to dramatically improve the efficiency and capabilities of assessment administration, scoring, and reporting.

Advances in the speed, capacity, and availability of computers allow applications that could only be imagined less than a generation ago.

College admissions and certification programs have led the way in using the new computer-based technology. The success of these pioneers has caused businesses ranging from the major commercial testing companies down to one-product start-ups to spend millions on assessment-related research and development. Transferring the emerging technologies fully into the K-12 arena seems an obvious next step.

Pushing beyond current testing limitations: What the future could look like

Computer-based assessment promises to both:

- make obsolete many of the shortcomings of current high-stakes, statewide assessment systems; and
- expand the capacity of such systems to measure rigorous standards in truly innovative ways.

State assessments are an essential tool in the student and school accountability model so prevalent in today's reform efforts. In large part because of the high stakes attached to their results, these new assessments have been subjected to intense scrutiny, and a number of problems have been identified, most having to do with the limits of the methodology. Proponents believe that new assessment technologies can break through these boundaries, solving many of the problems associated with traditional large-scale examinations.

Computer-supported scoring models have met or exceeded the accuracy of human raters across a range of content areas.

Broadly speaking, criticism of current large-scale assessment systems falls into three categories: logistics, content/methodologies, and value.

Logistics. Because they usually result from a series of compromises driven by limited time, resources, and methodologies, statewide assessment systems are far from ideal. For example, while schools would greatly benefit from detailed diagnostic information on each student, the amount of testing time and associated cost required to accomplish this on a statewide basis would be prohibitive using current methodologies. So while today's assessments provide schools with important programmatic data, more information would be needed for schools to develop individual student assistance plans. Equally desirable would be for schools to obtain results in a timely manner following administration of the assessment. Yet states that test in the spring and whose systems include direct writing or open-response items rarely receive the results before the end of the school year and often don't receive them until the following fall. Equally daunting is the cost and challenge of printing, delivering, and tracking the amount of paper used in high-stakes, statewide assessment systems.

Computer-based assessment provides answers to all of these problems. Using adaptive software, more detailed content can be tested in a shorter period of time. Results can be instantaneous, even for programs that require student writing. New breakthroughs in artificial intelligence and other models allow computer scoring of essays in a fraction of the time currently required. Computer-supported scoring models have met or exceeded the accuracy of human raters across a range of content areas. Computer administered tests require no printing and shipping of test booklets, a major expense for testing programs. When students respond online, no answer forms have to be shipped either, which cuts costs immediately. In addition, local staff time is no longer needed to process vast amounts of paper.

Content/Methodologies. The rigorous, "world-class" standards that have been adopted by many states stress students' ability to "know and do." However, with their heavy reliance on multiple-choice items (with some limited constructed-response items), many state tests emphasize more of the knowing than the doing. The new technologies open the door to using a range of methodologies not currently available, enabling not only the assessment of higher-level standards, but in formats potentially

more engaging than our current test stimuli. Use of CD-ROMS, for example, will allow students to respond to authentic "real-world" scenarios designed to assess problem solving, reasoning, and other higher-order skills, in addition to academic content. Simulation software already in existence for training programs are equally applicable in an assessment context. Graphing software can completely change the nature of mathematics tests, allowing students to develop and compare alternative solutions to challenging, real-world problems. Procedures such as "knowledge mapping," allow students to draw relationships among many concepts and then be scored against several varying "expert" maps. Students can create vastly different maps and still earn high scores because it is the reasoning behind the relationships that is being tested, not any single so-called "right' answer.

New methodologies hold out the promise of great improvements in the quality of data, leading to greater precision and increased validity.

Value. As noted above, computer-based assessment results are more immediately available to teachers than the results from current testing systems. This timeliness allows immediate intervention, whether it be remediation or enrichment, assignment to summer school, or anything else indicated by the results. More importantly, the *type* of information available from computer-based assessments promises to be much more valuable. Computer-adaptive testing (CAT) techniques, like those described earlier, zoom in on the ability level of each individual, allowing more reliable assessment with fewer items. Current systems require that students attempt every item even though, for some, the early items are too easy and for others, the final items are too difficult. Based on how a student responds to certain early items in an assessment, CAT models quickly identify that student's ability level and, thereafter, only present items that fit into that range. From an instructional perspective, the more "authentic" computer-based assessment methodologies can be readily integrated into classroom practices, allowing more of the curriculum-embedded features that were promised by proponents of performance assessments.

The new methodologies hold out the promise of great improvements in the quality of data, leading to greater precision and increased validity. Decreases in measurement error will be beneficial for the complex school accountability formulas in place in many states. Automated, adaptive assessments enable real-time equating, both vertical and horizontal, a feature currently lacking in both state and commercial testing programs.

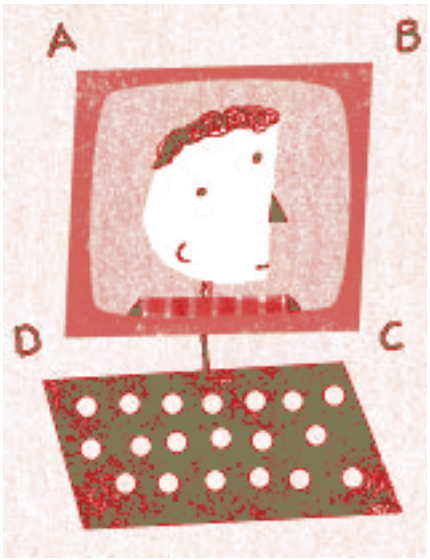
What stands in the way: Questions in need of answers

Given the notable advantages described above, the question is not *if*, but *when*, all high-stakes state testing programs will attempt to incorporate the emerging technologies. But standing between technology's promise of significantly more effective assessment and the fulfillment of that promise are a number of equally significant barriers. What follows is a discussion of the impediments that must be eliminated before technology can reinvent high-stakes assessment.

More Logistics. Unfortunately, use of the new technologies will not solve all existing logistics concerns, and it may even create some new ones. When a pencil point breaks, a student can simply sharpen it or switch to a new one. The problem of a computer crash is much more significant and often

time consuming to resolve. At the very least, the assessment delivery tool could be decommissioned in the middle of the test. There is also the potential that an entire testing session, along with the responses of all networked students, could be lost. Backup procedures are essential, both in terms of safely storing student responses and of having alternative means to administer the test.

Because schools differ tremendously in the hardware and software they have, system incompatibility is quite common whenever a new product line is introduced. In developing assessment systems that rely on technology, states need either to make sure the test software works on several different platforms or to require and support a single platform for all schools. Both options can be quite costly; the latter could exacerbate pockets of computer shortages and inequities.



The test administration process has two basic components: format of test stimulus (e.g., test booklets) and format of student response (e.g., student answer sheets). Currently, most programs use paper for both. Some programs have begun delivering tests on computer but continue to have students respond on a paper form. This limited approach is safer than relying entirely on computers because it reduces the likelihood of a computer crash, but it does not take full advantage of the technology. Paper responses must then be shipped and scanned, causing the delays and costs described earlier.

Delivery of materials and training of administrators remain necessary steps with computer-based assessment. Countless decisions must be made in the assessment administration process: Should files be stored on diskettes and sent by mail or should files be sent via the Internet? Will school officials at each site know how to load or access files and ensure uniform assessment conditions? Might certain software features need to be disabled (e.g., a spelling or grammar checker for a writing test), and does anyone in the school know how to do so? School personnel have become reasonably proficient in administering pencil-and-paper assessments. While computer-based assessment procedures are ultimately less difficult to implement, states must not underestimate the training needs in the early years of any new programs.

Scheduling is the final hurdle to overcome. Few schools have sufficient computer resources to test all students at once. How long it took before all students were tested would depend entirely on a school's resources, but it's conceivable that it could take from several days to several weeks. In such cases, *security*, *equivalence*, and *access* issues would arise. Each is discussed below.

Security. All current testing programs fear the copy machine. Complex security procedures are now standard practice in most states, driven in part by the high stakes now associated with statewide assessment. That said, security breaches in paper-based assessment systems, while not uncommon, tend at least to be localized and confinable. In contrast, with new digital technologies, a simple push of a button could send "secure" test forms literally around the world. While secure information is claimed to be "read only," stories of identity theft and hackers into the Pentagon missile systems should cause state test directors to wonder how safe any so-called secure files could ever be.

Another security concern would exist at each administration site. With the necessary extension of the testing window, students could more easily obtain advanced information on actual test items from those who took the test earlier. Computer labs would, of course, need to be supervised. But equally important, states might need to develop several forms of tests, drawing from much larger item banks than currently exist. However, few states have the resources to develop sufficient item banks for security purposes — or to take advantage of the type of CAT models described earlier. Unfortunately, those states that do develop large item banks would, in doing so, exacerbate the *equivalence* problem discussed next.

At the most basic level, we need to know that a test item presented "on screen" measures the same knowledge and skills as its paper counterpart.

Equivalence. Several significant concerns about the equivalence of the test administration setting, content, and results need to be allayed before states should jump fully ahead into computer-based assessment. Equivalence is important for a number of reasons, not least of which is the probability that, due to resource issues, many states might need to roll out a new technology-based assessment system in phases. In this scenario, some students in the state would be assessed with the new technology while others were assessed using traditional paper-based methodology.

Ensuring *administration* equivalence between a computer-based assessment system and the traditional pencil-and-paper system will require more research. Among the questions still needing to be answered: Are the number and timing of breaks needed during the assessment comparable from one medium to the other? How many test items should be included in any given session? Should tests be administered in one day or spread across more days? Should items be presented one at a time on the computer screen or should students be allowed to scroll across all items? Finally, for any of these questions, how might the answers differ according to students' age, sex, race, ethnicity, language, or disability?

Content equivalence would also need to be addressed. At the most basic level, we need to know that a test item presented "on screen" measures the same knowledge and skills as its paper counterpart. And what should be done about such features as a spelling checker, which is routinely available in word-processing packages? (Even making dictionaries available would require much more action and time from students than is required for using a computer's spelling feature). In the end, we must be able to say that content is identical across all delivery avenues.

Additionally, while computer-based assessment is expected to expand the types of standards that can be efficiently measured in large-scale assessment, states need to be sure they don't lose the ability to measure basic skills. The issues here closely parallel concerns raised for many years about the use of calculators on mathematics tests. The increased use of calculators led to math items typically being classified as calculator-neutral or calculator-enhanced, depending on the degree and type of calculation they require. Should future items be similarly classified according to how they would be affected by mode of delivery (e.g., technology-neutral)? This question would be especially important for the many states likely to begin their move toward computer-based assessment with mixed-delivery models, in which some students would receive paper-and-pencil tests while others received computer-based tests.

Finally, there is the issue of *results* equivalency. In the current high-stakes environment, any

perception (or reality) that one mode of assessment delivery or response results in higher scores than another will undermine the fairness of the system. In CAT designs, for example, *all* students can actually be taking different tests; in this scenario, aggregating assessment results for high-stakes accountability purposes could be difficult both to do and to explain. Even if we could develop formulas to statistically adjust various formats to make them equivalent, might the public not view this as a manipulation of scores to reach predetermined achievement goals or hide performance disparities? If all students took the test online, it would still be important to equate results across different years and formats lest we lose the ability to track trends, reward gains, and address the needs of those who continue to lag behind.

Access. The largest hurdle to realizing the promise of computer-based assessment is access, an issue that bumps right up against the inequities of funding and resource allocation evident throughout the public education system.

The first access factor is the need to assess using a mode of instruction the student commonly experiences. Take the content of the assessment. If the first time students encountered that content online was during the actual test administration, the results would not be valid. Just as curriculum and instruction must incorporate the content standards to properly prepare students for tests, so, too, must students be familiar with the actual technology used for assessment if such assessments are to be fair measures of student achievement.

Interestingly, a major impetus for many states to explore the new technologies has been the desire to accommodate special populations, as required by the federal Individual with Disabilities Education Act and other state and local statutes and regulations. English language learners, Title I students, and urban and rural students, among others, must all share the same comfort level if the emerging methods are to be judged fair and defensible.

Yet not all students have equal access to technology, either for instruction or assessment purposes. Nor are all teachers equally comfortable incorporating technology into their practice. Exacerbating the problem are the vast differences in students' access to computers outside of school. Many homes do not have a computer, let alone Internet connection. To the extent these disparities exist across socioeconomic, ethnic, and racial lines, performance differences on computer-based assessments would likely reflect or even increase the current achievement gap. States must determine how to provide equal (or at least sufficient) access to technology for both instruction and assessment before computer-based assessments can be considered a viable alternative to the status quo. Given the inability to do so over the past 20 years with calculators, a much less costly item, serious questions remain as to when equal access to technology might be achieved, permitting widespread implementation of innovative, emerging technologies. Also, with computer technology, as with calculators, attention must be paid to yet another type of equivalency: the power of the particular technology available. Currently, testing officials worry about students who work with four-function calculators having to compete with those whose calculators include the latest graphing capabilities. The potential for disparity is that much greater when dealing with varying hardware capacity and software packages.

Conclusion

Some revolutions are inevitable. Computer use is so central to today's society that it seems only a matter of time before computer-based assessment defines the test-administration process. States must plan *now* for that eventuality because, while the end of the road may already be determined, our ability to

transport all segments of society safely and fairly to that point is by no means guaranteed. Moving ahead without addressing the questions raised in this brief will almost certainly result in flawed, unfair, invalid (and most likely, illegal) assessment systems. Failure to move at all, however, threatens to deprive education of the many promises of computer-based assessment, most particularly, increased content coverage, decreased timelines, more defensible technical characteristics, and more authentic approaches to instruction and assessment.

WestEd, a nonprofit research, development, and service agency, works with education and other communities to promote excellence, achieve equity, and improve learning for children, youth, and adults. While WestEd serves the states of Arizona, California, Nevada, and Utah as one of the nation's Regional Educational Laboratories, our agency's work extends throughout the United States and abroad. It has 16 offices nationwide, from Washington and Boston to Arizona, Southern California, and its headquarters in San Francisco.

For more information about WestEd, visit our Web site: WestEd.org, call 415.565.3000 or, toll-free, (1.877) 4-WestEd, or write:

WestEd
730 Harrison Street
San Francisco, CA 94107-1242

This report was produced in whole or in part with funds from the Office of Educational Research and Improvement, U.S. Department of Education, under contract #ED-01-CO-0012. Its contents do not necessarily reflect the views or policies of the Department of Education.

WestEd®
730 Harrison Street
San Francisco
California 94107-1242

Address service requested