

Title: Considerations in statewide implementation of computer-adaptive testing

Date: June 2008

Question: Which states are using computer adaptive testing, and what are the possible benefits of this approach and considerations and lessons learned from statewide implementation?

Response:

As technology has become increasingly more prevalent and accessible for use in student assessment, several states are adopting innovative test delivery models to collect, analyze, and report student-level data. Among these models, computer-adaptive testing (CAT) is garnering attention as a means to assess student performance and progress with fewer items and less testing time. However, to build a CAT-based state assessment system that complies with the provisions of the No Child Left Behind (NCLB) Act, states will have to surmount a number of logistical, infrastructural, regulatory, and financial hurdles, something only Oregon has done successfully.

Computer-Based Testing Versus Computer-Adaptive Testing

Many states are introducing computer administration of their statewide assessments alongside or in lieu of paper-and-pencil (P&P) administration. Computer-based testing, or CBT, refers to assessment delivered via computer. CBT can take many forms ranging from tests that deliver items linearly on the computer just as on paper to computer-adaptive testing (CAT), in which items are presented based on each student's responses to previous items. Most often, CBT refers to tests in which all students take the same form (or, more likely, one of a small number of forms), whereas in CAT, there is no fixed form, since the sequence of items on each test varies according to the student's responses. Proponents of CBT and CAT claim that compared to P&P administration, both computer modes allow for faster score reports, inexpensive presentation of high-resolution images, introduction of innovative item types, certain efficiencies in test administration, enhanced security, and greater access and use of technology resources available at schools (CTB/McGraw-Hill, 2004; Drasgow, 2002; Klein & Hamilton, 1999). Above and beyond these benefits, CAT holds the potential for more customized assessment with items that are tailored to students' ability levels and identification of students' skills and weaknesses using fewer test items and requiring less time (Shorr, 2002). On the other hand, a CAT system requires more items to be developed and sophisticated computer algorithms to select items.

States Using CBT and CAT

Although adaptive tests have been implemented by many school districts for both summative and formative purposes, there are only a handful of states that have implemented statewide computer-adaptive testing. According to *Education Week's* "Technology Counts," 27 states are implementing computer-based assessment in 2008 (*Education Week*, 2008). As of 2005, the latest date for which comprehensive data are available, states delivering at least some portion of their statewide assessments via adaptive tests include the District of Columbia, Idaho, Maryland, Oregon, South Dakota, and Virginia.¹ North Carolina had pilot-tested computer-adaptive tests between 2000 and 2002.

Oregon's Success and Idaho's Challenge

The contrasting experiences of Oregon and Idaho regarding federal review of CAT are instructive for a state considering this assessment approach. While both states developed their own web-based computer testing system with input from educators statewide and with the help of a contractor, Oregon's test was shown to meet NCLB criteria of item alignment with grade-level content standards while Idaho's test initially did not. In early 2006, the U.S. Department of Education found that Oregon needed to provide additional documentation showing that the adaptive testing for its general assessment, the Oregon Assessment of Knowledge and Skills (OAKS), tested grade-level standards and provided comparable achievement level results to the P&P administration (U.S. Department of Education, 2006). Though the Oregon items are selected based on student responses to items from earlier in the test, all items assess grade-level standards. Furthermore, Oregon demonstrated that all students would be tested on the full range of content specified in the state assessment blueprint. By December 2007, the general assessment met all federal requirements.

On the other hand, when the Idaho Standards Achievement Test (ISAT) adapted items based on student responses, the items selected could be "off grade level," sometimes several grade levels above or below the student's grade. As a result, in late 2005, the U.S. Department of Education concluded that the ISAT was not aligned with state content standards. Idaho discontinued its assessment contract with the initial assessment developer. After switching vendors, Idaho implemented a test with a core set of items that could be used for federal accountability purposes and an adaptive set that could help to more accurately estimate students' ability levels.

Other States' Experiences with CAT

The Maryland Functional Testing Program – Computer Adaptive Test (MFTP-CAT) in reading and mathematics was a set of online computer-adaptive tests administered during the 2002/03 and 2003/04 school years. These tests were available only for students retaking tests (*Education Week*, 2008). Tests could be administered on either a PC or

¹ Education Week does not cite the District of Columbia among states with statewide adaptive testing in 2005 (though the District is included in the publication's 2003 report and did have adaptive testing in 2005). In addition, Mississippi is cited as administering statewide adaptive testing. In fact, Mississippi's online Student Progress Monitoring System (SPMS) has never been adaptive (L. Kramer, personal communication).

Macintosh platform. The test would stop when a standard error criterion was reached or a maximum number of items had been administered (35 for the math test, 30 for the reading test). Content balancing occurred over the course of the test via the item selection algorithm. Students had limited review of items that they had answered previously (NCS Pearson and Maryland State Department of Education, 2002).

In Virginia, adaptive testing is designated for at-risk students. The online computer-adaptive Algebra Readiness Diagnostic Test (ARDT) is administered to students identified as at risk of failing the Algebra I end-of-course test. With a pre- and post-test design, the ARDT assesses the knowledge and skills of the Mathematics Standards of Learning in Algebra I for students in grades 3 through 8 and measures the impact of the Algebra Readiness Initiative (ARI), an intervention model aimed at preparing students to be successful in Algebra (<http://vardt.starttest.com>). Currently, 95 percent of schools in Virginia are participating in the ARI and most districts utilize the ARDT when they accept funding for the intervention (<http://www.aypf.org/forumbriefs/2008/fb04112008.htm>). In South Dakota, the State Department of Education has contracted with a computerized-testing company to develop and administer the Performance Series, an online standards-based adaptive test designed for students in grades 2 through 12. As part of the Dakota Assessment of Content Standards (DACS), the test is criterion-referenced and aligned to South Dakota standards (<http://www.edweek.org/ew/articles/2001/03/07/25sd.h20.html>).

North Carolina and the District of Columbia have piloted computer-adaptive tests designed to meet the unique needs of students with disabilities. From 2000 to 2002, North Carolina implemented an adaptive version of its state reading and mathematics assessments for special education students (Russo, 2002; L. Kramer, personal communication). Designed to give disabled students access to test items and scale score results, the North Carolina Computerized Adaptive Testing System (NCCATS) was an online version of a paper-and-pencil test that had some success, but according to the Department of Public Instruction, was never ready for full implementation (http://www.sreb.org/programs/EdTech/pubs/PDF/Online_Testing.pdf). The District of Columbia has also piloted CAT for students in alternative and special education programs. The district has used the Measures of Academic Progress (MAP) for diagnostic and monitoring purposes to provide instructional data on students as they enter and progress through alternative programs.

A tabular summary of states with computer-adaptive testing programs (including discontinued programs) is presented in Table 1.

Table 1. States with Computer-Adaptive Testing Programs (including discontinued programs)

State	Name of test	Part of state accountability?	Subject	Formative / summative	Eligible students
District of Columbia	Measures of Academic Progress (MAP)	No	Reading, Mathematics	Formative	Students in alternative and special education programs
Idaho	Idaho Standards Achievement Test (ISAT)	Yes (For purposes of determining AYP, only the grade-level tests are used.)	Reading, Language Usage, and Mathematics (grades 3–8, 10); Science (grades 5, 7, and 10)	Formative and summative	All students in grades 3–8, and 10
Maryland (Discontinued 2004)	Maryland Functional Testing Program – Computer Adaptive Testing (MFTP-CAT)	No	Reading, Mathematics	Formative	Retakes only
North Carolina (Discontinued 2002, although research continued on it until 2005)	North Carolina Computerized Adaptive Testing System (NCCATS)	No	Math, Reading	Summative	Students with disabilities (“gap” students) in grades 3–8 and grade 10
Oregon	Oregon Assessment of Knowledge and Skills (OAKS)	Yes	Math, Reading, Science, and Social Sciences	Summative	Grades 3–8, HS
South Dakota	Performance Series, which is part of the DACS (Dakota Assessment of Content Standards)	No	Mathematics, Reading, Language Arts, and Science	Formative	Grades 2–12
Virginia	Algebra Readiness Diagnostic Testing (ARDT)	No	Algebra	Summative (pre/post) to measure the impact of the Algebra Readiness Initiative (ARI)	Grades 5–8 students who participate in ARI are also required to participate in the ARDT

Possible Benefits of CAT

At the core, the proposed benefits of CAT boil down to two essential elements, efficiency and diagnosis. When students take test items that are close in difficulty to their ability levels, fewer items are needed to establish a precise estimate of student ability. Fixed-form

assessments need to include items at a range of difficulty levels, to obtain a precise estimate of the abilities of all examinees. This requires more items on a test form and longer testing times, and in this sense CAT is more efficient than fixed-form assessment. Several studies that have documented the comparability of CAT to P&P administration have found CAT to be a valid and reliable means of measuring student progress (see Klein & Hamilton, 1999 for review). Well-designed CATs can provide more accurate scores over a wide range of abilities than traditional tests (Meijer & Nering, 1999). It has also been found to measure student progress more precisely. For example, the Delaware Statewide Academic Growth Assessment pilot project found adaptive tests to be better able to identify students' academic growth than grade-level tests (Hoff, 2007). Administrators have also found that CAT reduces test length by half (Kosty et al., 2006; Lilley et al., 2004) and testing time by more than half, while maintaining the same level of validity and reliability (Rudner, 1998). Shorter testing times lead to less demand on hardware access and capacity as well as reduced fatigue in test takers (Rudner, 1998). In addition, the dynamic nature of CAT can provide greater student interaction and motivation than is afforded by traditional CBT (Lilley et al., 2004).

By selecting future items based on previous responses, CAT can more quickly diagnose gaps in student understanding. Fixed-form assessments, by contrast, do not diagnose learning gaps efficiently. At the district level, CAT is often seen as a way to improve diagnostic testing and explore skills and abilities not directly assessed on current tests (Russo, 2002). Researchers have noted that adaptive testing can be an efficient strategy to collect meaningful information that serves both accountability and instructional purposes (e.g., Buckendahl, 2005).

Technical, Practical, and Regulatory Considerations

Before implementing a statewide system of CAT, however, policymakers need to think through a host of technical, practical, and regulatory considerations. For example, CAT requires an expanded item pool, and rules for replenishing that item pool are generally more complicated than those of a fixed-form assessment. This can have substantial cost implications. In addition, the infrastructure required for either statewide CAT or CBT is extensive, and for most states significant expenditures on equipment must be made before any statewide computer-delivered assessment system is feasible. Furthermore, there are a number of requirements of the assessment and administration design that must be met to satisfy requirements of NCLB.

Technical Considerations

The technical issues associated with computer-adaptive tests are many and complex. Large banks of calibrated items have to be created and continually updated to ensure item and test security (Kosty et al., 2006; Meijer & Nering, 1999). In fact, Way et al. (2006), citing findings from Stocking (1994), report that “an item pool equal to about 12 times the length of a fixed-length CAT was adequate for a variety of content areas and test structures” (p. 6). This is several times larger than alternative designs, which may consist of as little as a single form and one or two backup forms that are used in the event of a test security breach. Building the larger CAT item pools can represent a significant cost—of time as well

as money—to the state. Developing the item selection algorithm, deciding when to end the test, ensuring content balance, and developing administration procedures (such as whether students may review and change their past responses) are other technical issues that are beyond the scope of this brief. (See Way et al., 2006 for a detailed review of technical issues.)

Practical Considerations

In addition to the challenge of adapting CATs to meet NCLB regulations (see next section), there are several logistical challenges that arise in implementing computer-based assessment. On the district level, documented challenges in implementing CAT include maintaining the correct computer program as technology shifts, increasing teacher use of the system, workstation incompatibility, problems with outdated browsers, network and bandwidth limitations, and the technological inexperience of teachers and local computer support staff (Russo, 2002). On the state level, there are concerns about districts' and schools' readiness to support new assessment systems that incorporate CAT as it initially involves more logistics and potential issues such as system incompatibility, delivery of materials, variation in students' familiarity with the format, training of administrators, and scheduling use of computer resources (Rabinowitz & Brandt, 2001). Ensuring the presence of adequate computer infrastructure across the state—and that it is working properly—is a challenge. When the computers supporting a CBT are not working, the same linear P&P form can be given instead; by contrast, the adaptive benefits are lost when the computers supporting a CAT are down.

Regulatory Considerations: CAT and NCLB

Given the detailed regulations set forth in the NCLB law, it is essential that states and districts first consider the purpose of using CAT—whether for diagnostic or accountability purposes or both. In addition, it is worthwhile to consider how computer-based testing may affect concerns about access and validity, as well as security and privacy (Russo, 2002). As noted in Oregon and Idaho's review process, the U.S. Department of Education requires states to measure student performance against the expectations for a student's grade level in order to fulfill the requirements of the No Child Left Behind law (*Education Week*, 2003). Computer-adaptive tests are often considered "out of level" because the range of items can include skills and content offered in higher and lower grades (Trotter, 2003). According to a CTB/McGraw-Hill (2004) report,

The U.S. Department of Education has determined that an assessment administered in a strictly computer-adaptive test (CAT) mode, in which items that assess standards from grades other than the one in which the student is enrolled are included in the test, does not comply with [the on-grade] requirement. Therefore, states could jeopardize receipt of funds under Title I of NCLB if this approach were adopted (pp. 3-4).

While this stipulation curtailed efforts in several states to introduce CAT as part of their assessment programs, it does not prohibit the use of CAT for NCLB accountability purposes *as long as items in the CAT measure relevant grade-level standards* (Way et al., 2006). In order to gain federal approval under the NCLB law, state education officials must show that computer-adaptive tests measure achievement standards that are aligned to the state's

grade-level content standards and produce evidence of the reliability, validity, and comparability of the CAT to standard P&P forms (U.S. Department of Education, 2007). While the NCLB law requires state tests to be aligned with state content standards, it is challenging to conduct alignment studies for an adaptive test where there is no fixed form (Kosty et al., 2006). According to Kosty et al. (2006), there is a need to establish criteria and standards to justify and support the use of CAT and to clearly demonstrate to policymakers and educational administrators the value added by such tests. Critics of the current law highlight the duplicative nature of administering CAT and traditional tests (*Washington Post*, 2007) and note that public awareness and federal funding for the development of CAT will remain limited as long as computer-adaptive tests are excluded from the federal law (Trotter, 2003).

Lessons Learned

While computer-adaptive testing has yet to be universally accepted in high-stakes, statewide assessment systems, there is great potential for CAT to improve both the systems of assessment and the quality of the information gathered (Rabinowitz & Brandt, 2001). As Rabinowitz (2005) states, “computer-adaptive methodology can, in the long term, address the issue states face of either having to increase the time and cost of assessment or risk insufficient reliability at key decision points” (p. 279). CAT requires large item banks that need to be continually refreshed, particularly important when CAT is used for high-stakes accountability purposes. The Internet affords an efficient and cost-effective mode of delivery for CATs, with central maintenance reducing the burden on individual school sites to install or modify products (Klein & Hamilton, 1999). However, issues of fairness and access must be addressed in the short term, given the diversity of technological resources across districts, schools, and classrooms (Klein & Hamilton, 1999; Rabinowitz, 2005).

Despite NCLB stipulations, several states at the forefront of CAT—including Oregon, Idaho, and South Dakota—have moved forward with incorporating CAT into their statewide assessment programs. Solutions include administering a fixed-form test for NCLB purposes followed by adaptive items that are based on responses to the fixed test or giving a separate CAT either following a paper assessment or in the beginning of the school year to be used as a diagnostic tool (CTB/McGraw-Hill, 2004). Following a two-year pilot project using Measures of Academic Progress (MAP), a CAT that was aligned to state standards, officials in Delaware are recommending two types of tests be implemented as part of the state assessment program: a comprehensive exam to use for accountability purposes and a diagnostic test to use for instructional purposes (Kepner, 2008). Similarly, students in Idaho now take grade-level exams, from which data for NCLB are generated, followed by a computer-adaptive set of items. South Dakota and Oregon have also made changes to their statewide assessment programs as a result of NCLB regulations focusing on aligning CAT items to state standards. Above all, it is important to recognize that incorporating CAT into statewide assessment systems takes time and is best implemented in steps.

Statewide implementation of computer-adaptive testing is relatively new, and its impact and implications are still being discovered. However, some findings emerge from research into CAT systems and from the experiences of states that have built them:

- Well-designed CAT systems can reduce testing time.
- Item pools supporting CAT systems need to be larger than those for other testing modes.
- The Internet holds promise for efficient and cost-effective CAT administration. The diversity of technological resources across districts, schools, and classrooms presents a challenge to standardized CAT administration.
- Although CAT systems are not prohibited by NCLB, adaptive tests must test students on content tied to grade-level standards and must cover the breadth of content specified in assessment blueprint.
- Testing students with a fixed-form test for NCLB purposes combined with adaptive items or a separate CAT may be a means to obtain the benefits of CAT while complying with federal legislation.

Advances in technology, improvements in infrastructure, and changes in legislation and regulatory guidance will all shape the cost, feasibility, and attractiveness of statewide computer-adaptive testing as states move into the new decade.

References

Note: Many of the sources in the bibliography are publishers of computer-adaptive tests. Due to the cost and lead-time necessary to develop the CAT platform and technology, it is difficult to find research studies from a party that has no financial interest in the success of CAT. This potential conflict of interest should be kept in mind when reviewing these sources.

Applying adaptive technology to diagnose student performance and progress. (2005). Scantron Corporation.

Buckendahl, C. W. (2005). *Challenges for instructionally supportive accountability tests*. Paper presented at the annual meeting of the Northern Rocky Mountain Educational Research Association, Jackson Hole, WY.

Computer-Based Testing – Issues and considerations. (2004). CTB/McGraw-Hill.

Drasgow, F. (2002). The work ahead: A psychometric infrastructure for computerized adaptive testing. In C. N. Mills, M. T. Potenza, J. J. Fremer, & W. C. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments*. Mahwah, NJ: Lawrence Erlbaum Associates.

Education Week. (2003). Pencils down: Technology's answer to testing (*Education Week*, "Technology Counts" special issue).

Education Week. (2008). Education counts: Research center. Retrieved October 28, 2008, from <http://edcounts.edweek.org/createtable/step1.php>

- Hoff, D. J. (2007). The latest news on the reauthorization of the No Child Left Behind Act. *Education Week*, retrieved November 4, 2008, from http://blogs.edweek.org/edweek/NCLB-ActII/2007/10/the_next_version_to_nclb.html
- Kepner, A. (2008). Online test adjusts to pupils: Pilot project demonstrates faster feedback at lower cost, educators say. *The News Journal*, January 9. <http://www.delawareonline.com/apps/pbcs.dll/article?AID=2008801090337>
- Klein, S. P., & Hamilton, L. (1999). *Large-scale testing: Current practices and new directions*. Santa Monica: RAND.
- Kosty, D., McBride, J., Poggio, J., Wise, L., & Way, D. (2006). *What's next in online testing*. Presentation from the 36th Annual National Conference on Large-Scale Assessment, San Francisco, CA. Retrieved November 4, 2008, from <http://www.ccsso.org/content/PDFs/Session38Way.pdf>
- Kramer, L. Mississippi Department of Education, Jackson, MS.
- Lilley, M., Barker, T., & Britton, C. (2004). The development of a software prototype for computer-adaptive testing. *Computers & Education*, 43, 109–123.
- Meijer, R. R., & Nering, M. L. (1999). Computerized adaptive testing: Overview and introduction. *Applied Psychological Measurement*, 23, 187–194.
- National Education Technology Plan (posted 2005, online). Washington, DC: U.S. Department of Education. Retrieved November 3, 2008, from <http://www.nationalEdTechPlan.org/>
- NCLB, testing and flexibility. (2007). *Washington Post*, November 27. Retrieved October 28, 2008, from http://www.publiceducation.org/nclb_articles/archive/20071127_NCLB.asp
- NCS Pearson and Maryland State Department of Education. (2002). Maryland Assessment Group Conference: Computer adaptive testing [PowerPoint presentation]. Retrieved November 21, 2008, from http://www.magonline.org/docs/MAG%20Conference%20Yoes_Arnold.ppt
- Overview of U.S. Department of Education's position on the use of computer-adaptive testing*. (2004). CTB/McGraw-Hill.
- Rabinowitz, S. (2005). The integration of secondary and post-secondary assessment systems: Cautionary concerns. In W. Camara & E. W. Kimmel (Eds.), *Choosing students higher education admissions tools for the 21st century*. Routledge.

- Rabinowitz, S., & Brandt, T. (2001). *Computer-based assessment: Can it deliver on its promise?* San Francisco: WestEd.
- Rudner, L. M. (1998). *An on-line, interactive, Computer Adaptive Testing tutorial*. Retrieved November 3, 2008, from <http://echo.edres.org:8080/scripts/cat/catdemo.htm>
- Russo, A. (2002, April). Mixing technology and testing: Computer-based assessments lend flexibility, quick turnaround and lower costs, supporters say. *School Administrator*.
- Shorr, P. W. (2002, Spring). A look at tools for assessment and accountability. *Administrator Magazine*.
- Stewart, A. K. (2008). Utah is hoping to avoid Internet-test pitfalls. *Deseret News* (Salt Lake City), Aug. 31, 2008. Retrieved November 3, 2008, from <http://www.deseretnews.com/article/1,5143,700255175,00.html>
- Technology Counts 2007: A digital decade. *Education Week on the Web*, March 29, 2007. Retrieved November 3, 2008, from www.edweek.org/go/tc07
- Trotter, A. (2003). A question of direction. *Education Week on the Web*, May 9, 2003.
- U.S. Department of Education. (2006, January 25). *Oregon assessment letter*. <http://www.ed.gov/admins/lead/account/nclbfinalassess/or.html>
- U.S. Department of Education. (2007). *Standards and assessments peer review guidance: Information and examples for meeting requirements of the No Child Left Behind Act of 2001*. Retrieved October 28, 2008, from www.ed.gov/policy/elsec/guid/saaprguidance.doc
- Way, W. D., Davis, L. L., & Fitzpatrick, S. (2006). Practical questions in introducing computerized adaptive testing for K-12 assessment. *Pearson Educational Measurement*.

Additional Resources

The National Assessment of Educational Progress is conducting research and field tests on technology-based assessment.

<http://nces.ed.gov/nationsreportcard/studies/tbaproject.asp>

The Education Commission of the States compiles useful articles and information about state activities related to computer-based testing. <http://www.ecs.org>

Testing Our Schools (PBS Website):

<http://www.pbs.org/wgbh/pages/frontline/shows/schools/>

This memorandum is one in a series of quick-turnaround responses to specific questions posed by educators and policymakers in the Western region (Arizona, California, Nevada, Utah), which is served by the Regional Educational Laboratory West (REL West) at WestEd. This memorandum was prepared by REL West under a contract with the U.S. Department of Education's Institute of Education Sciences (IES), Contract ED-06-CO-0014, administered by WestEd. Its content does not necessarily reflect the views or policies of IES or the U.S. Department of Education nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.