Computerized Adaptive Testing

How CAT May Be Utilized in the Next Generation of Assessments

A Report for the
North Carolina State Board of Education

July 16, 2010
STATE BOARD OF EDUCATION

The guiding mission of the North Carolina State Board of Education is that every public school student will graduate from high school, globally competitive for work and postsecondary education and prepared for life in the 21st Century.

WILLIAM C. HARRISON
Chairman :: Fayetteville

REGINALD KENAN
Rose Hill

JOHN A. TATE III
Charlotte

WAYNE MCDEVITT
Vice Chair :: Asheville

KEVIN D. HOWELL
Raleigh

ROBERT “TOM” SPEED
Boone

WALTER DALTON
Lieutenant Governor :: Rutherfordton

SHIRLEY E. HARRIS
Troy

MELISSA E. BARTLETT
Roxboro

JANET COWELL
State Treasurer :: Raleigh

CHRISTINE J. GREENE
High Point

PATRICIA N. WILLOUGHBY
Raleigh

NC DEPARTMENT OF PUBLIC INSTRUCTION
June St. Clair Atkinson, Ed.D., State Superintendent
301 N. Wilmington Street :: Raleigh, North Carolina 27601-2825

In compliance with federal law, NC Public Schools administers all state-operated educational programs, employment activities and admissions without discrimination because of race, religion, national or ethnic origin, color, age, military service, disability, or gender, except where exemption is appropriate and allowed by law.

Inquiries or complaints regarding discrimination issues should be directed to:

Dr. Rebecca Garland, Chief Academic Officer :: Academic Services and Instructional Support
6368 Mail Service Center, Raleigh, NC 27699-6368 :: Telephone: (919) 807-3200 :: Fax: (919) 807-4065
Visit us on the Web:: www.ncpublicschools.org
Executive Summary

Computerized adaptive testing involves computer-based or online delivery of an assessment that is tailored to individual students’ ability levels. A student’s interaction with each item impacts the delivery of the next item. Specifically, a student who answers an item correctly will receive a harder item, and a student who answers an item incorrectly will receive an easier item.

Computerized adaptive testing is different from the current system of assessments in that assessments currently target precise measurement around the proficiency level, thus sacrificing precision for students whose ability is far below or far above the cut score. Assessments under an adaptive framework are tailored for each student’s ability. A number of testing programs have adopted adaptive testing as their model for assessment. These programs include the Armed Services Vocational Aptitude Battery (ASVAB), the Graduate Record Examination (GRE), the Graduate Management Admissions Test (GMAT), and the Oregon Kindergarten–12th grade assessment program.

Computerized adaptive testing for state assessments has all the benefits of an online delivery system for assessments but goes beyond those advantages to include a number of other benefits. As stated previously, one major advantage of computerized adaptive testing is the ability to produce reliable scores for students along all points of the ability continuum, particularly those students who struggle to meet proficiency and those students, like Academically or Intellectually Gifted (AIG) students, who far exceed levels of proficiency in a grade. It is also possible to produce more reliable subscores depending on the defined purpose of the assessment. Additional opportunities for improvement of the current assessment system that result from delivery of computerized adaptive assessments relate to: (1) the administration of assessments, such as reduced costs for printing, shipping, and storing paper-based test booklets; (2) the utility of the assessment system, such as the usefulness and timeliness of the data resulting from end-of-year assessments; and (3) the appropriateness of assessments, such as the use of authentic items aligned to 21st-century skills and knowledge.

The major challenge in moving toward an assessment system using technology either to deliver assessments via computer or to deliver computerized adaptive tests is the initial development cost for the system and infrastructure necessary to support such a system. The state will address this challenge by creating an implementation plan for moving to an online delivery system for the next generation of assessments.

The North Carolina Department of Public Instruction (NCDPI) has a history of producing computerized adaptive tests. A program was produced prior to the No Child Left Behind Act of 2001 that allowed students to be assessed on grade-level content that was suited to their ability, regardless of the grade level they were assigned. Due to the philosophy of NCCATS being in
conflict with federal regulations under No Child Left Behind, the project was shelved. There appears to be a new opportunity to move the direction of computerized adaptive testing, albeit with clear parameters dividing assessments by grade level. State consortia working to build common assessments have also announced preparations to pursue computerized adaptive tests designed to measure the Common Core State Standards in English language arts and mathematics. As a result, the NCDPI has created a timeline for implementation of field tests of a newly redesigned computer adaptive testing system. This initiative falls in line with the State Board of Education’s desire to pursue innovative solutions and assessment systems that utilize technology. Computerized adaptive testing represents an effort to align assessments with next generation innovations by utilizing available technology and to design an assessment system that is appropriate for all students.
Introduction
The mission of educators in North Carolina is to prepare every student to graduate from high school globally competitive for work, post-secondary education, and life in the 21st century. As part of this mission, the state is committed to providing an assessment system that not only informs instruction and evaluates knowledge and skills, but also uses technology in a way that is relevant to 21st-century learners. The North Carolina Department of Public Instruction (NCDPI) is involved in a number of initiatives that will help meet these goals. The agency’s Framework for Change (2008) laid the groundwork for the Accountability and Curriculum Reform Effort (ACRE) which encompasses the overall vision of moving toward the next generation of assessments. Another initiative in which the agency is involved is the national state-led movement toward a common core of content standards in English language arts and mathematics (Common Core State Standards, CCSSs) and common assessments across state consortia. Both efforts focus on assessments that are delivered using 21st-century technology and assessments that include innovative, technology-enhanced items.

The NCDPI recently published a paper outlining the benefits and challenges of online assessments as well as the state’s readiness to deliver assessments online (NCDPI, 2010). In summary, the benefits included both benefits to the state and local education agencies (LEAs) in terms of limiting the resources required for assessment as well as benefits to students in terms of the opportunity to utilize more authentic measurement techniques. The challenges relate to the technological capacity of the state and LEAs to support online assessment delivery. At the outset, the investment in preparing LEAs will be onerous but the long-term leveling of costs, and even reduction in required resources, combined with the other benefits make online assessment a valuable tool.

The NCDPI’s commitment to 21st-century learning and assessment environments gives a unique opportunity to explore all the possible innovations that online assessment allows. One option that becomes available in an online assessment environment is adaptive testing. The remainder of this document will outline the essential points for consideration when moving to a computerized adaptive assessment system.

What Is Computerized Adaptive Testing and Why Is It Valuable?
In simplest terms, computerized adaptive testing (CAT) involves a computer-based (or online) assessment that adjusts to an examinee’s ability level. In essence, a CAT can be tailored to an individual student’s ability. This type of assessment is in contrast to the tests that North Carolina currently uses, which are paper-based or computer-based linear tests that are comprised of predetermined items with a fixed length. In a CAT, a series of moderately difficult items are administered to a student to establish a baseline estimate of the student’s ability. Using that
baseline estimate, a computer algorithm determines the next item to be administered to the student that will maximize the accuracy of the score. In its simplest form, if a student answers an item incorrectly during a computerized adaptive test, the next item delivered to the student is an easier item. If the student answers the item correctly, the next item delivered to the student is a harder item.

Computerized adaptive testing is a tool that the North Carolina Testing Program has pursued in the past and could be restored to meet the principles of the North Carolina State Board of Education (SBE) surrounding preparation of students for the 21st century. CAT has enormous value in ensuring all students are provided with appropriately challenging assessments that produce highly reliable scores. By targeting the items administered to each students’ ability level, CAT ensures that every student who is tested is challenged by each item administered but is not discouraged by facing items that are too difficult or by facing items that are too easy (Wainer, 2000). In addition, students across the ability range obtain more equally reliable scores. This approach is in contrast to the current assessments, which target the most precise measurement at grade-level proficiency and do not provide individualized assessments.

The typical approach to summative testing is to administer the same, or very similar, assessments to all students within a very short time frame. By assessing all students in a grade with a single form of the assessment, it is clear that a broad range of ability is being tested, and the implication is all students can be appropriately measured with the same sets of items. Since many students’ ability can be reliably assessed using items of moderate difficulty, assessments are typically weighted with larger numbers of items of moderate difficulty and fewer items in the higher and lower range of difficulty. The drawback of this approach is some students, such as low-performing students or students who are academically gifted, are faced with items that are either too difficult or too easy for them; therefore, little information about their true skills and ability is realized. CAT would maximize the reliability of scores for all students by presenting more items that match a student's ability.

**CAT for Summative or Diagnostic Purposes**

CAT can be used, like any assessment, for various purposes. The delivery of the assessment and scores reported would depend on the intended use of the assessment. A CAT used for diagnostic purposes might be delivered continually as part of a state-based benchmarking tool. An assessment delivered more frequently would span a more narrow range of content and might provide more diagnostic information about students’ knowledge and skills related to a specific topic. Using CAT in this manner would allow teachers to have clear and immediate data to use to customize instruction to meet the needs of individual students. A summative assessment, on the other hand, would span a broader range of content and would be less diagnostic but would still have the advantage of being a precise measure for all students across the ability continuum.
Advantages of CAT
CAT systems can be appropriate for either benchmark or summative purposes, but the extent of the advantages and challenges is dependent on the purpose of the assessment. Advantages may be more or less relevant depending on the purpose of the assessment. The list presented here generally applies to a CAT used for summative purposes. Appendix A contains a full description of the advantages. The advantages of CAT can be grouped into four general categories:

- Administration,
- Technicality,
- Utility, and
- Appropriateness.

<table>
<thead>
<tr>
<th>Administration</th>
<th>Reduced testing time</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Reduced burden on LEA staff</td>
</tr>
<tr>
<td></td>
<td>Increased flexibility in test scheduling</td>
</tr>
<tr>
<td></td>
<td>Improved test security</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Technicality</th>
<th>Increased overall score reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>More reliable subscore reporting</td>
</tr>
<tr>
<td></td>
<td>Simple administration of experimental items</td>
</tr>
<tr>
<td></td>
<td>Increased quality control for poorly performing items</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Utility</th>
<th>Increased information available to test users</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Increased availability of diagnostic data</td>
</tr>
<tr>
<td></td>
<td>Increased flexibility in use of innovative items</td>
</tr>
<tr>
<td></td>
<td>Positive impact on student motivation</td>
</tr>
<tr>
<td></td>
<td>Aligned assessment system components</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Appropriateness</th>
<th>Increased flexibility in accommodations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Customized assessments</td>
</tr>
</tbody>
</table>

CATs provide many administrative advantages. CATs with no constraints on the content of the items can be substantially **shorter than traditional tests** because items are administered such that information about a student’s true ability is maximized with every item, thus reducing the time required of students to spend on testing. There is also a reduction in the time LEA staff expends on sorting and handling test materials, hence, **reducing the burden on LEA staff** by reducing the time required of them to prepare for, administer, and conduct test administrations. Removing these administrative concerns **improves test security** by limiting the opportunities for individuals to handle the test materials. Similarly, LEA staff might also benefit from an **increased flexibility in scheduling** test administrations. Smaller groups of students can take the assessment over the course of multiple testing sessions without causing concern for test security because virtually **every assessment is unique**.
**Technicality**

CAT offers the ability to maximize certain technical characteristics of an assessment system. As mentioned previously, criterion-referenced tests typically target the most precise measurement at the cut score for proficiency, meaning many items are used that align to the ability required to reach proficiency. Conversely, CAT allows for **precise measurement of all students**, regardless of where they fall on the ability continuum. Research in Oregon showed that their CAT assessment reduced standard errors in the extremes of the ability distribution. That is, these students were being measured more reliably than they would have been if given the paper and pencil version of the assessment.⁴ The notion of increased overall reliability can also be extended to subscore reporting for diagnostic purposes. A CAT used for benchmark purposes could produce highly reliable scores for all students and could be narrowly focused on specific content units such that results immediately inform teachers how to adjust instruction to deal with areas of opportunities where students need more development. Other technical advantages include the ease with which experimental items are field-tested in an online environment and, conversely, the ease with which poorly performing items can be identified and removed from the item pool. **Experimental items are easily embedded** into the online environment, thus keeping motivation high while field-testing new items, making their statistics more reliable and valid. **Poorly performing items can be removed immediately** by having real-time evaluations of test results.

**Utility**

Test developers are engaged in a constant effort to balance the needs of administrators and teachers with the technical adequacy of information produced by a summative assessment system attempting to serve multiple purposes. Typically, test scores are only valid for one purpose. For example, a summative assessment intended to evaluate the delivery of instruction across the state cannot also be used for student diagnostic purposes, hence, the need to balance the desires of test users with the appropriateness of their usage of data. Again, because a CAT system also benefits from the opportunities afforded by online assessments, CAT systems provide **more information** than has typically been available in paper-based systems. Because information on student response time by item will be available, data users and researchers might gain insight into the motivation of students to answer certain types of items. Data can also be captured on the types of items that students flag for review, how many times they change their answers, or if certain test-taking strategies (such as highlighting answer choices or editing tools for writing-based items) are used during the test administration. This information would likely lend insight into students’ thought processes during problem solving. Additional **diagnostic information** might be available when using the assessment as a benchmark tool.

---

¹ Criterion-referenced tests refer to an interpretation of students’ scores relative to a specified expectation of mastery of content. In interpreting students’ scores, there is no reference to performance of other individuals in determining students’ competence in the subject matter.

² Standard errors were reduced by 3 to 6 scale score points for students in the tails of the ability distribution. See attached excerpt from CAT presentation (Alpert, 2010).
Because CAT uses an online delivery system, all the benefits of online assessment delivery systems also apply to CAT, particularly the ability to include innovative item types that are more aligned to 21st-century learning than what is possible using a paper-based assessment system. Students are increasingly at ease with technology and are using technology in their classrooms more frequently. Assessments of learning should align with 21st-century classrooms by also utilizing technology that is relevant to students, which would likely increase their motivation to attend to the tasks required of them.

Finally, immediacy of feedback to stakeholders (e.g., principals, teachers, parents) is of the utmost concern for an assessment system. Currently, North Carolina is one of very few states that provides same-day scoring. Score turnaround for multiple-choice items in a CAT system can be maintained and potentially enhanced depending on policy decisions. If desired, scores could be reported to students immediately upon completion of an assessment; no scanning of student answer sheets would be necessary.

**Appropriateness**
The adaptive nature of a CAT avoids the “one-size-fits-all” approach for which summative assessments are currently criticized. Students will be faced with test items that are appropriately challenging, not test items that are too easy or too difficult. There is added flexibility in an online testing system which provides a favorable platform for providing accommodations that prove challenging and/or costly in a paper-based system. Universal Design principles — those assessment design principles that attempt to include all students in a single design — can be applied to an online assessment in the same way they are in a paper-based system.3 However, the online environment provides increased opportunities for the types of accommodations available and the application of Universal Design principles.

**Challenges of CAT**
Many of the challenges of moving toward a CAT system are challenges faced by simply moving to an online system. The advantages outweigh the challenges in number, but the challenges are, nonetheless, substantial. See Appendix B for a full list and description of the challenges. There are a number of challenges North Carolina would face in moving to a CAT system, some of which the state would face in moving to any online assessment system. These challenges include:

- Costs,
- Technology,
- Shifting testing paradigm,
- Field testing,
- Score reporting, and
- Item exposure

3 See NCEO Synthesis Report No. 44 (2002) for more information on Universal Design.
The most substantial challenge facing any state moving toward a CAT system is the cost for such a system. The **costs** stem from both the assessment itself, including the development and maintenance of an online delivery system, but also from the need to build capacity statewide to enable districts to successfully administer online CATs. Another cost is the need for high quantities of test items. However, in an established assessment system like North Carolina’s, this burden is not as excessive or immediate because robust item pools already exist and the maintenance of these item pools is ongoing.

Technological challenges will also be significant but can be overcome if funding is available. The NCDPI report titled *Transitioning to Online Assessment in North Carolina* (2010) outlining the state’s readiness to transition to online assessments will serve as direction for where resources are most needed.

Moving to a CAT will be a major **shift in the testing paradigm** of North Carolina. First, a major assessment literacy effort will be necessary to define what an adaptive system is and the advantages it provides to multiple test users. Related to this shifting paradigm will be changes to score reporting that require stakeholders to understand scores can no longer be reported as the raw number of items students answered correctly because students will likely receive different “forms” of the tests of differing lengths. Moreover, because of the past experience with adaptive testing as an accommodation in North Carolina and the numerous technological challenges in the field and from the state level, it will be imperative the NCDPI educate stakeholders as to how these challenges have been addressed and ensuring those challenges will no longer be an obstacle.

Lastly, **item exposure and security** will need to be monitored closely. The selection algorithms that are used to identify which items are administered to students often use the most informative items to begin the CAT, causing those items to become overexposed. If flexibility is given in scheduling test administrations there must be some control over having too many students exposed to the same items. Overexposure threatens test security but also threatens the accuracy of the adaptive algorithm if the content of these starting items is exposed and students become familiar with how to answer correctly. The adaptive nature of the assessment may then be inaccurate for a student if he/she truly does not have the ability to answer these items correctly.

**Are Other Testing Programs Using Computerized Adaptive Testing?**

A number of testing programs use adaptive testing formats. For example, the Graduate Record Examination (GRE) and Graduate Management Admissions Test (GMAT) all utilize adaptive testing for all or part of their exams. However, there are currently very few states using CAT for high stakes assessment (i.e., federal or state accountability). Only one state, Oregon, has been approved to use CAT for federal accountability purposes. Their assessment, Oregon’s Assessment of Knowledge & Skills (OAKS), was fully approved by the United States Department of Education.
(USED) in March 2009. Key features of the OAKS are that all general assessments are administered as adaptive tests, the length of the testing window is extensive, students are given multiple opportunities to test, and the highest score for a student is used for all reporting and accountability purposes. Specifically, the testing window for the OAKS opens in October of an academic year and stays open until the end of May. Students are allowed three opportunities to test. The highest of the three scores is used for student, school, and state accountability. The system currently utilizes only multiple-choice items in order to ensure immediate score reporting. Full approval of Oregon’s system is evidence that a CAT can be a successful and accepted tool for evaluating student performance for high stakes accountability purposes. Because of its success, Oregon’s system is serving as the model for the work of a multi-state common assessment consortium called the SMARTER Balanced Assessment Consortium (SBAC) that intends to build CATs aligned to Common Core Standards. North Carolina is currently participating as a governing member of this consortium.

Other states have made ventures into the use of CAT, but the implementation of those assessments are at different stages, and the purposes for which those assessments are developed and used vary. Phillips (2009) reported 10 states, including North Carolina, which have or have had some form of CAT already in existence or under development. Further communication with representatives from each state indicated there is quite a bit of variability in the use of CAT and the purposes for which results are used. Beginning in 2010–11, Hawaii will join Oregon in moving its entire assessment system into an adaptive environment. The Hawaii Department of Education used the Oregon model as the foundation for designing their new assessment system. The additional 8 states reported in Phillips (2009) have either investigated and abandoned the use of CATs (e.g., North Carolina) or do not use their CAT systems for summative federal accountability purposes (e.g., South Dakota and Idaho).

**North Carolina’s History with CAT**

North Carolina is a state that has experience with CAT. In 2000–01, the NCDPI implemented an online CAT as an accommodation for students with disabilities taking the reading or mathematics End-of-Grade tests in grade 3–8 or the High School Comprehensive Test at grade 10. With the inception of NCLB in 2001, the North Carolina CAT (called NCCATS) was discontinued. Although implementation of the NCCATS faced significant obstacles in terms of the costs related to its development and the limited capacity of both the state to support school districts and school districts to provide sufficient quantities of computers and consistent connectivity and bandwidth, the insurmountable barrier at the time was the resistance of the federal government to allow adaptive testing for accountability purposes. NCLB required assessments used for AYP reporting to be comprised only of items aligned to a single grade-level. The NCCATS, like many adaptive assessments, included items that spanned grade-levels in an effort to identify items that were best-suited to individual students. As mentioned previously, the USED has recently approved
Oregon’s adaptive assessment system as the state was able to prove that their CAT assessments were adaptive within grade levels and did not include off grade-level items (USED, 2009).

In order to move forward, investment in revamping and revising the existing NCCATS system is necessary. This effort will include a timeline for pilot testing and long-term implementation. As part of the initiative, clear decisions about the purpose of such a system need to be outlined so the necessary technical and administrative parameters are built into the system during development.

**Implementing Computerized Adaptive Testing**

There are a number of practical matters a state should contemplate when planning for a CAT as a component of the assessment system (Way, 2005). These practical matters are laid out above and are generally associated with psychometrics, item development, and the assessment delivery system. North Carolina must also evaluate the technological infrastructure of schools and districts and provide solutions in instances where the infrastructure will not support an online testing environment. There are a number of steps that should be taken to prepare for an assessment of this kind. These steps are outlined in the graphic below. This graphic will serve as an implementation guide for delivery of a CAT in North Carolina over the next three years. By building on North Carolina’s previous CAT system, the state could significantly shorten the development timeline, and a new system could be in place for a limited number of content areas in approximately three years.

**Timeline for Implementation of CAT**

<table>
<thead>
<tr>
<th>2010</th>
<th>2011</th>
<th>2012</th>
<th>2013</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spring Fall</td>
<td>Spring Fall</td>
<td>Spring Fall</td>
<td>Spring Fall</td>
</tr>
<tr>
<td><strong>Prepare for CAT System</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Evaluation of State Readiness</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Item Development/System Prototyping</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Determination of Delivery Criteria</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Usability Study</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Field Testing</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Operational CAT</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Conclusions**

It is clear that a number of improvements to the current assessment system are possible by moving to an online CAT assessment system. These include administrative advantages that reduce testing time and burden on LEA staff, technical advantages that ensure reliable measurement of all
students, system utility advantages that produce more information for test users, and system appropriateness advantages that allow increased flexibility in customizing assessments to all students' needs. Additionally, laying a foundation for a CAT-based assessment system aligns closely with the work of the SMARTER Balanced Consortium of which North Carolina is a governing state. In a new decade, with increased technology available to schools and a new perspective on how to implement a CAT that would meet NCLB requirements, it is appropriate to begin investigating how a valuable tool like computerized adaptive testing might be used to meet the demands of the next generation of assessments.
References


## Appendix A: Advantages of CAT

### Administrative Advantages

<table>
<thead>
<tr>
<th>Advantage</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reduced testing time</td>
<td>CAT with no constraints can be substantially shorter than fixed-length tests because items are administered such that information about a student’s true ability is maximized with every item administered. It is unlikely the state testing program would be able to reap the rewards of this advantage because of federal requirements to cover a wide range of content on end-of-year assessments.</td>
</tr>
<tr>
<td>Reduced burden on LEA staff</td>
<td>The burden of receiving test materials, sorting and handing out test materials, collecting and scanning test materials, and shipping any required materials back to the state is eliminated.</td>
</tr>
<tr>
<td>Increased flexibility in scheduling</td>
<td>A more secure testing environment allows for multiple test administrations to accommodate individual school schedules. Smaller groups of students can take the assessment at multiple test administrations rather than all students taking the assessment at once.</td>
</tr>
<tr>
<td>Improved test security</td>
<td>Tests are no longer able to be copied or inappropriately retained as study guides as each assessment is different for each student.</td>
</tr>
</tbody>
</table>

### Technical Advantages

<table>
<thead>
<tr>
<th>Advantage</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Increased overall reliability</td>
<td>When tests are criterion referenced, it is typical for the highest reliability to be around the cut score between proficient and not proficient. With a CAT, there is increased reliability of scores for students above and below the cut score because items are targeted at individual students' abilities such that score information is more reliable. In one example, the difference between paper &amp; pencil and CAT reliability indices ranges between 3 and 6 standardized units at the ends of the ability distribution with less error reported across the scale with CAT.</td>
</tr>
<tr>
<td>Reliable subscore reporting</td>
<td>Related to diagnostic feedback, CAT can serve an additional purpose by constraining benchmark assessments to be of a sufficient length to provide reliable subscores for units of the Essential Standards. Reliability is inherently improved by targeting items to a student’s ability but adding the additional constraint that unit-based assessments be of a certain length ensures reliable subscore reporting.</td>
</tr>
</tbody>
</table>

---

4 See attached excerpt from CAT presentation (Alpert, 2010).
<table>
<thead>
<tr>
<th>Simplicity in field-testing new items</th>
<th>Field-test items are embedded within the operational assessment.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Removal of poorly performing items</td>
<td>Although items are currently field-tested and screened before being placed on operational assessments, one cannot predict if an item will behave unexpectedly under operational conditions. An online environment provides the opportunity for real-time monitoring of assessment data. Therefore, an item that behaves unexpectedly can be identified during assessments and removed from the pool so that testers during the same administration are not exposed to the item. In the current system, those items are removed after the test administration, if necessary, but are still administered to all students before it is determined whether a concern exists.</td>
</tr>
</tbody>
</table>

**System Utility**

| More information provided to educators | Teachers – Receive information specific to each student representing mastered knowledge and skills  
Principals – Access to detailed information about the performance of students across classrooms which may help identify successful instructional strategies that can be applied to areas of opportunity  
State – Because a larger number of items spanning a content area are potentially administered, state officials can get a wider view of the delivery of the entire span of Essential Standards, Clarifying Objectives, and finer points within those content standards.  
Data Users – Online delivery systems can be developed to capture response times, changes to answers, number of items flagged for review, etc. that might provide insight into students’ thought processes. These types of data aid in instruction and future test development. |
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Diagnostic</td>
<td>CATs used for benchmarking purposes can be constrained to contain items from certain units of the Essential Standards that relate to pacing in the classroom but, within that constraint, items are still targeted at a student’s ability such that information can be gleaned about where students may have weaknesses.</td>
</tr>
</tbody>
</table>
### Innovative items & motivation

21st-century learners are using technology in their daily lives. Having an assessment system that utilizes technology makes assessment more relevant to students. Additionally, the online testing environment provides opportunities to include innovative items that cannot be included on a paper and pencil test. Having an assessment system that is relevant to students increases the likelihood that students will be engaged in the assessment.

### Immediate scoring

CATs are delivered via computer (either housed on local machines or centrally delivered) using items that have been previously field-tested and scaled such that immediate scoring can be provided through the online environment. There is no need/delay to scan answer sheets when using an online administration.

### Alignment

An online summative assessment is aligned with the vision of benchmark assessments. Having both the benchmark and summative assessments delivered in an online environment allows for increased opportunities to align the two assessments in terms of item content and scoring.

### Advantages for System Appropriateness

<table>
<thead>
<tr>
<th>Additional accommodations</th>
<th>Online testing environments provide greater opportunity to incorporate Universal Design principles and standardized accommodations that are challenging and costly in a paper and pencil environment.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avoids one-size-fits-all</td>
<td>Tests are specifically targeted at each individual student’s ability.</td>
</tr>
</tbody>
</table>
## Appendix B: Challenges of CAT

<table>
<thead>
<tr>
<th>Challenges of CAT</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Costs</strong></td>
</tr>
<tr>
<td><strong>Technology</strong></td>
</tr>
<tr>
<td><strong>Public perception</strong></td>
</tr>
<tr>
<td><strong>Field-testing</strong></td>
</tr>
<tr>
<td><strong>Score reporting</strong></td>
</tr>
<tr>
<td><strong>Item exposure</strong></td>
</tr>
</tbody>
</table>

---

⁵ See detailed report on the feasibility of shifting the South Carolina assessment system to a CAT (Data Recognition Corporation, 2007). This report is available upon request.
Attachment

The following attachment is an excerpt from a presentation given by Tony Alpert, Director of Assessment at the Oregon Department of Education, at a meeting on Best Practices for State Assessment sponsored by the National Academy of Science.
Evidence for adaptive testing

Standard Error of Measurement by Scale Score and Assessment Mode
Grade 8 Mathematics