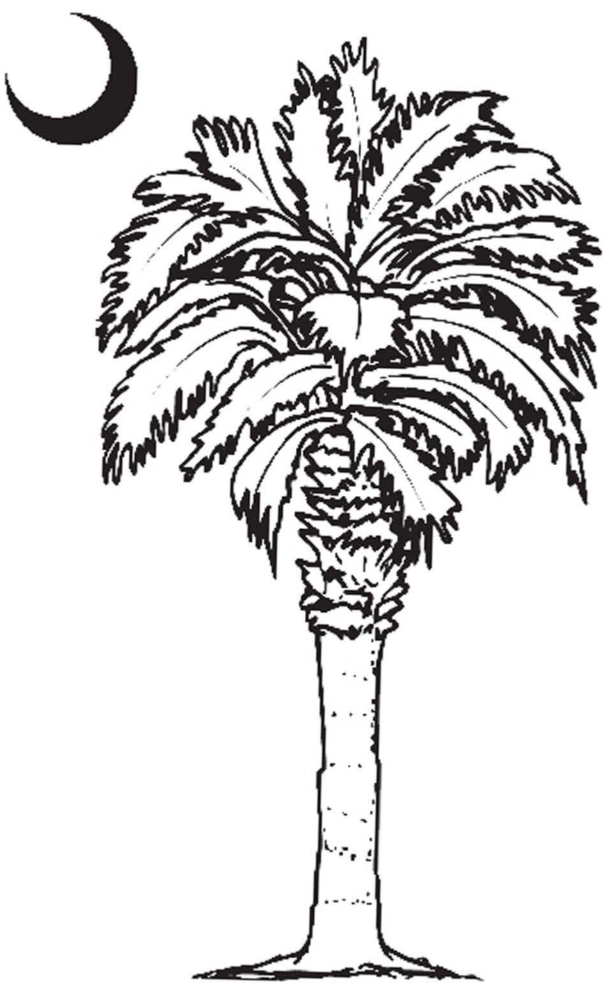
Study on the Feasibility and Cost of Converting the State Assessment Program to a Computer-Based or Computer-Adaptive Format

FINAL REPORT



June 25, 2007

Table of Contents

ELECTRONIC TESTING: TERMS AND INTRODUCTORY CONCEPTS

Computers are revolutionizing education in general and testing specifically. This study focuses on test administration, although many related issues such as design, content, accommodations, and reporting must also be carefully reviewed and considered. To avoid confusion in an emerging field with many similar-sounding terms, the study begins with the definition of key terms. The definitions apply to this document only. Note that other authors and other documents, including those listed in the references, may make different distinctions.

Electronic testing (eTesting) refers to any assessment presented to the student via a computer screen and with which the student interacts via a keyboard, mouse, or other pointing device. It may or may not involve the Internet and World Wide Web. It may or may not involve branching decisions to tailor the test to the student. *Computerized testing* is considered to be synonymous with eTesting and will be used interchangeably.

Computerized adaptive testing (CAT) refers to an electronic test that uses branching to tailor the test to the student's level of proficiency. In general, no two students will receive the same set of items. This topic will be discussed in detail in the section of this report entitled: *Computerized Adaptive Testing*.

Computer-based testing (CBT) will be restricted here to refer to the administration of fixed forms, in contrast to CAT. Generally, all students will take the same items. However, it does include the case of multiple versions of the form with the same items in different orders (i.e., scrambled forms). It could also include several parallel fixed forms. *Parallel fixed forms* means all forms are built prior to testing using the same content and statistical specifications but each form will contain different items. Both scrambled forms and parallel forms are used to enhance security.

COMPONENT 1 – DESCRIPTION OF STATE ASSESSMENT PROGRAMS WHICH ARE COMPUTER-BASED OR COMPUTER-ADAPTIVE

This component will address the following:

- Results from an e-mail questionnaire sent to all states regarding CBT/CAT initiatives
- Summary of the *Education Week*'s *Technology Counts 2007: A Digital Decade* findings
- Recent media reports on statewide CBT/CAT testing
- Overview of accommodations that can be and are being offered by states via computer-delivery of tests
- Overview of keyboarding and other technology state standards

State E-mail Questionnaire

In February 2007, the following questionnaire was sent to all state departments of education. All fifty state departments responded.

1. Does your state offer any statewide computer-delivered testing? If no, please skip to question 11.
2. What grade and subject-area tests are offered via computer?
3. Is testing via computer voluntary or mandatory?
    - If voluntary, about what percent of testing takes place via computer?
4. Approximately how many computer-delivered tests were administered in your state last year (i.e., number of tests delivered via computer)?

5. Are the computer-delivered tests "fixed" or "scrambled" forms, or computer-adaptive?

Please indicate one of the following:

- Fixed (i.e., every form is "fixed" with the items appearing in the same order)
- Scrambled (i.e., the same items are given to every student, but they are "scrambled" to appear at different locations on the test)
- Computer-adaptive (i.e., the test items are chosen and presented to a student based on that student's response to previous items)
- A combination

6. Do the computer-delivered tests include <u>short-constructed</u> responses? (yes or no)

- If yes, in what grades and subjects are these offered?
- How are these responses scored? (choose the best option below)
- "by hand" (i.e., responses are delivered to human scorers for scoring); if so, by what means (e.g., scanning, direct delivery via online)?
- "artificial intelligence" (i.e., a computer program scores the responses)
- a combination of the two scoring methods are used (please explain)

7. Do the computer-delivered tests include <u>extended-writing</u> responses? (yes or no)

- If yes, in what grades and subjects are these offered?
- How are these responses scored? (choose the best option below)
- "by hand" (i.e., responses are delivered to human scorers for scoring); if so, by what means (e.g., scanning, direct delivery via online)?
- "artificial intelligence" (i.e., a computer program scores the responses)
- a combination of the two scoring methods are used (please explain)

8. How quickly are computer-delivered test scores made available to the student, school, district, and state (i.e., what is the reporting turnaround time for these different groups)?

- Student results:
- School results:
- District results:
- Statewide results:

9. Which vendor(s) are providing your test engine(s) for your computer-delivered testing?

10. Is there anything else you would like to share from your state's experience in implementing computer-delivered testing that might inform the State of South Carolina's Computer-Based/Computer-Adaptive Feasibility Study?

11. If your state does not currently offer any statewide computer-delivered testing, is your state planning (or has conducted) any pilots or feasibility studies of computer-delivered testing? (if yes, please explain)

Summary of Findings from E-mail Questionnaires

STATE OF THE STATES CONDUCTING CBT/CAT

Relatively few states have implemented across-the-board computerized testing for grades 3–8. These include Oregon, Kansas, Virginia, Wyoming, Utah, and Idaho, with only Idaho, Wyoming, and Oregon making critical NCLB/AYP online tests mandatory for almost all students.

Of the latter:

- **Idaho** notes it "has been happy with our computer-delivered tests. Change is always hard, and some school districts were very apprehensive when we stared the computer-delivered tests. At this point however, there would not be one school district that would wish to go back to paper and pencil."

- **Wyoming** states, "…Students and schools really like the online testing, but it does impact the use of computers and tech resources during the testing window. Wyoming districts are tech ready and early adopters. They were willing to experience the glitches in order to have more rapid reporting and better control of test security."

Most other states which are moving their testing programs into the online arena have focused or will initially focus their computerized testing initiatives on non-NCLB grades/subjects (i.e., end-of-course tests, writing, geography), as they are less "high stakes." For example: West Virginia has mandated online writing assessments for grades 7 and 10; Indiana and South Carolina offer high school end-of-course exams voluntarily via computer; Oklahoma has started with mandatory grade 7 geography test online; Kentucky tests its Special Populations students online; and North Carolina mandates its Computer Skills exam be taken via computer. North Carolina also offers its end-of-course tests online; however, only the physics test is mandated as an online assessment because it includes "innovative item types" that cannot be replicated via paper/pencil.

STATE-LEVEL INTEREST IN, AND CONCERNS WITH, IMPLEMENTING CBT/CAT

Responses from many states that are not currently involved in statewide CBT or CAT indicate significant interest in transitioning to online assessments. Several states have conducted, or are planning for, pilots and other small-scale studies of computer-based testing. Others, such as Delaware, are including online testing plans in recent contracts or upcoming Requests for Proposals.

Of the states providing insights into why they are not currently offering CBT or offering only limited CBT, the following sorts of challenges were mentioned: capacity, funding, technical support, and the need to ensure a "full-service" vendor for both paper/pencil and CBT. Specifically:

- **Vermont** notes: "We are very interested in computer-delivered testing, but our issue is capacity. We still have many schools with dial-up connections, and insufficient terminals to make web-based testing feasible."
- **Massachusetts** indicates that the state is considering options for online testing in 2008; however, state funding is not expected to cover the requisite costs for this initiative. Thus, "online testing plans are now on hold."
- **Alaska** notes that connectivity and the number of computers is a concern. They are starting with formative assessments.
- **Colorado** notes they are conducted pilots on writing, "but this is a long-term capacity issue."
- **Connecticut** states, "We have plans to convert our test into a computer-ready version in the future. Piloting will give us part of the answer to the feasibility question. There are certainly questions of access and load that are not yet answered in Connecticut."

- **Indiana** (which offers high school end-of-course tests online) notes: "Special challenges include finding a vendor that is a one-stop shop for delivering, scoring, reporting, and psychometric services for CBT and P/P assessments (simultaneously)."
- **Oklahoma** (which has begun CBT testing with Geography grade 7) noted during the Expert Panel Meeting that, with regard to rural districts, "It's not the hardware issues, it's the technology staffing issues, no tech support in these rural districts…[SC should] think about this in terms of training."

All state responses have been sorted into two categories: YES, we are doing CBT or CAT at a statewide level; and NO, we are not doing CBT or CAT at a statewide level. Summary matrices of responses are located below in Tables 1 and 2. Complete, verbatim responses from each state are included at the end of Component 1.

Additional feedback and insights from states that have implemented CBT or CAT initiatives are found in Component 12's discussion of the Expert Panel meeting held in Columbia, South Carolina on March 28, 2007.

Table 1.1 Summary Matrix of States Answering "Yes" They Offer CBT/CAT

State	Offers CBT	Grades/Subjects	Constructed Response/ Extended Response	Testing Window	Report Turnaround	Voluntary/ Mandatory	Vendor (for test engine)	*Technology Counts* (2007) Student: Instructional Computer Ratio (based on 2006 data)
Florida	Yes	Grade 10, retakers only, Reading/Math	No	Multiple 5-day windows	6 weeks, all data	V	CTB/ McGraw Hill	3.5
Georgia	Yes	HS EOC tests (8)	No	Multiple 5–6-week windows	5 days	V	Pearson	3.8
Idaho	Yes	Grades 2–10 Reading/Math/ Language Usage; Grades 5, 7, 10 Science	No	4/16–5/11	Immediate—preliminary, final 1 week	M	CAL	3.6
Indiana	Yes	HS EOC tests (4)	Yes, including essay (human scoring and AI)	4/30–5/31 plus fall/ winter windows	24 hrs. from close of student test, after 10 days of testing	Winter M; Spring V	Questar	3.4
Kansas	Yes	Grades 3–8, 11 Reading/Math; Grades 6, 8, 10 Social Studies	No	2/1–4/16	Student results within 24 hrs.; unofficial district results mid- to late-May	V	N/A: Center for Educational Testing and Evaluation (CETE), University of Kansas	2.7

Table 1 (continued). Summary Matrix of States Answering "Yes" They Offer CBT/CAT

State	Offers CBT	Grades/Subjects	Constructed Response/ Extended Response	Testing Window	Report Turnaround	Voluntary/ Mandatory	Vendor (for test engine)	*Technology Counts* (2007) Student: Instructional Computer Ratio (based on 2006 data)
Kentucky	Yes	Grades 3–8, 10–11 (special populations) Reading/Math/ Science/Social Studies/Writing/ and others	Yes, prompts (human scoring)	4/23–5/4	Reported with P/P—approx. August 1	V	2007 CTB & eCollege; 2008 Measured Progress	3.8
Maryland	Yes	Grades 5, 8 Science	Yes, CR (human scorers, piloting AI for second read)		N/A 2007 = first year, standard-setting, etc. must occur before reporting. Plans for summer reporting after 2007.	V	Pearson	4.8
Minnesota	Yes	Grades 3–8, 11 Math for ELL (operational 2007); Grades 5, 8 Science High School (operational 2008)	CR for science (human scorers)	4/16–5/4 (according. to news reports, extended)	2–3 months for all reporting	V	Pearson	3.8

Table 1 (continued). Summary Matrix of States Answering "Yes" They Offer CBT/CAT

State	Offers CBT	Grades/Subjects	Constructed Response/ Extended Response	Testing Window	Report Turnaround	Voluntary/ Mandatory	Vendor (for test engine)	*Technology Counts* (2007) Student: Instructional Computer Ratio (based on 2006 data)
Mississippi	Yes	High School EOC tests (4), retesters only	No	Multiple dates for single-day testing and single-day make-ups	3 weeks for pass/fail rosters; 5–6 weeks for student score reports	M for retests	Harcourt	4.6
North Carolina	Yes	Computer Skills (grad. requirement) and High School EOCs (7)	"In a sense, for CS and Physics—pattern-matching or captured keystrokes, scored dichotomously"	Computer Skills, multiple testing windows, EOCs, the last week of course.	For student-level reports, generally the next day.	M for Computer Skills, V for EOCs except Physics (due to innovative item types)	N/A: State partners with NC State University for test engine	4.0
Oklahoma	Yes	Grade 7 Geography, with plans for all grades (3–8)/subjects	No	4/10–4/2; Four testing windows planned for EOI tests	Raw scores available immediately; school results available about 1 week after close of testing window	M	ITS	3.9

Table 1 (continued). Summary Matrix of States Answering "Yes" They Offer CBT/CAT

State	Offers CBT	Grades/Subjects	Constructed Response/ Extended Response	Testing Window	Report Turnaround	Voluntary/ Mandatory	Vendor (for test engine)	*Technology Counts* (2007) Student: Instructional Computer Ratio (based on 2006 data)
Oregon	Yes	Grades 3–12, Reading/Math/ Science/English Language Proficiency, K–12	MC (plus some CRs for English Language Proficiency exams—human and AI scoring)	September–May (Note: Used Paper/ Pencil Spring 2007 due to CBT/CAT delivery problems)	Student results, immediately.	M	Vantage-academic; Language Learning Solutions—English Proficiency	4.5
South Carolina	Yes	High School EOCs (4)	No		Student results within 36 hours; school summaries within 4 to 5 weeks.	V	Pearson	3.7
Texas	Yes	Grades 7–10, and Exit Level Retest—all subjects tested; plans for grades 3–6 Reading Proficiency Test in English II (grades 2–12 only online); High School EOC tests (6).	Yes	2/20–2/22 (Writing 4, 7 = MC + composition Reading 9 = MC + Open-ended ELA = MC + Open-ended + composition—Human Scoring. 4/16–4/20 (MC-only tests)	Same as for P/P—2 weeks after MC administration.	V	Pearson	3.4

Table 1 (continued). Summary Matrix of States Answering "Yes" They Offer CBT/CAT

State	Offers CBT	Grades/Subjects	Constructed Response/ Extended Response	Testing Window	Report Turnaround	Voluntary/ Mandatory	Vendor (for test engine)	*Technology Counts* (2007) Student: Instructional Computer Ratio (based on 2006 data)
Utah	Yes	Grades 4–12 Science; Grades 3–11 ELA; Grades 3–12 Math	No	The last 5 weeks before the end of the school year.	Class reports available in 48 hours	V	Pearson	
Virginia	Yes	Grades 3–8 Reading/Math; Grades 3, 5, 8 Science/History; High School EOCs (12)	No	Statewide Spring window is 4/16–6/29. Divisions must select one of three windows for their Grades 3–8 tests in the spring. Options for grades 3–8 windows are about 3 weeks each.	Complicated with VAs pre- and post-equating. Pre-equated forms: almost immediately; non-pre-equated forms; non-equated forms, approx. 2 weeks.	V	Pearson	3.1
West Virginia	Yes	Grades 7, 10 Writing	ER only (scored by AI)	Grade 7 3/12–3/23; Grade 10 2/26–3/9	All reports within 60 days	M	Vantage for AI scoring	3.0
Wyoming	Yes	Grades 3–11, Reading/Math/ Science	Yes, starting in 2008, 3–11 (human scorers)	3/26–4/27	MC = second week of testing window; all results 4 weeks after close of testing window.	M	Harcourt	2.4

Table 1.2 Summary Matrix of States Answering "No" They Do Not Offer CBT/CAT

State	Offers CBT	Non-CBT/CAT States with Pilots/Studies Conducted or Planned; or Plans to Include in RFP	Vendor (for test engine) If Indicated for Piloting/Studies	*Technology Counts* (2007) Student: Instructional Computer Ratio (based on 2006 data)
Alabama	No			4.7
Alaska	Yes Formative Only		CAL	2.7
Arizona	No	Yes	Pearson	4.5
Arkansas	No			3.6
California	No	Yes		5.1
Colorado	No	Yes		4.2
Connecticut	No	Yes	Measurement Inc.	4.0
Delaware	No	Yes		4.5
Hawaii	No			4.9
Illinois	No	Yes		3.8
Iowa	No			3.3
Louisiana	No	Yes	Pacific Metrics	4.1
Maine	No	Yes		1.9
Massachusetts	No	Yes		3.6
Michigan	No	Yes		3.7
Missouri	No		N/A: MCCE is part of the University of Central Missouri.	3.5
Montana	No			2.7
Nebraska	No			3.0
Nevada	No	Yes	Probably their current contractor, Measured Progress	5.0
New Hampshire	No	Yes		5.2
New Jersey	No			3.7
New Mexico	No			3.6

Table 2 (continued). Summary Matrix of States Answering "No" They Do Not Offer CBT/CAT

State	Offers CBT	Non-CBT/CAT States with Pilots/Studies Conducted or Planned; or Plans to Include in RFP	Vendor (for test engine) If Indicated for Piloting/Studies	*Technology Counts* (2007) Student: Instructional Computer Ratio (based on 2006 data)
New York	No			4.2
North Dakota	No			2.9
Ohio	No			3.5
Pennsylvania	No			3.5
Rhode Island	No			5.0
South Dakota	Yes Formative Only		Scantron	1.8
Tennessee	No	Yes	Pearson/ETS	4.0
Vermont	No			3.5
Washington	No			3.9
Wisconsin	No			3.2

Additional Information on CBT/CAT Gathered Post-Questionnaire

As several states noted in their responses to the questionnaire, "glitches" can and do occur with online testing. Since compiling the questionnaire responses, there have been widely-reported "glitches" with three statewide online testing programs this spring. While problems with all online testing programs can occur (and should be expected, just as occasional shipping/delivery problems can occur with paper/pencil tests), problems are more often localized, short-lived, and/or identified during piloting. The particular instances below have been widely-reported in the press, perhaps because of the high-stakes nature of the tests involved and the number of students/districts affected.

Specifically:

- Oregon's online testing program [Technology Enhanced Student Assessment (TESA), math, science, and reading, grades 3-12—with a projected 1.2 million tests delivered online this year] experienced problems with its online testing vendor this spring and opted to revert to paper/pencil testing until next fall. (See, e.g., Ann Williams, "It's back to pencil, paper for state tests," *The Register-Guard* (Eugene, OR) May 7, 2007; Oregon State Department of Education Website)

- Minnesota implemented its first operational online administration of the Mathematics Test for English Language Learners (MTELL) for grades 3-8 and 11 this spring. There were issues with download time for audio; the testing window had to be extended. (See, e.g., Associated Press, "Online math test stymies ESL students," *Post-Bulletin* (Rochester, MN) April 27, 2007.)

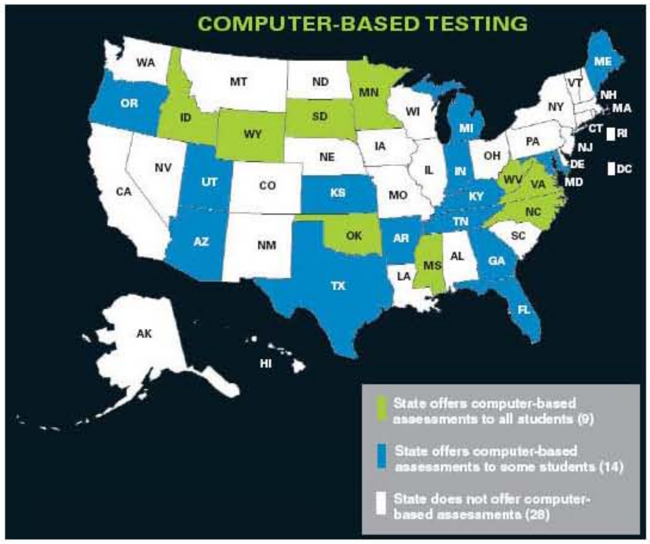
- Virginia, which has been testing NCLB tests online since 2001, experienced online delivery problems this spring. Vendor server problems required approximately 9,500 students to retest. (See, e.g., Gold, M. "Firm Pledges to Fix Online Exam Glitches," *Washington Post* (Washington, D.C.) June 5, 2007.)

EDUCATION WEEK'S TECHNOLOGY COUNTS 2007: A DIGITAL DECADE

Although the questionnaire sent to state departments of education should provide the most up-to-date information, additional data regarding CBT/CAT testing initiatives, as well as broader state-level technology issues, has been compiled by *Education Week*'s *Technology Counts 2007: A Digital Decade.* The findings in this national report are summarized below. (The full report is available in hard copy and online, "Detailed State Reports" are available only online at: http://www.edweek.org/ew/toc/2007/03/29/index.html)

Technology Counts states, "the number of states that offer computerized statewide assessments is relatively small, with 14 states making that opportunity available on a limited basis, such as within certain districts, or for students retaking pencil-and-paper tests. And only nine states offer computer-based testing to all students." Those nine states make at least one statewide assessment available on the computer to all students.

NOTE: 1) *Technology Counts* does not include more specific information on which tests are given, whether they are mandatory, and which students have access to those tests; 2) Two of the nine states listed by *Technology Counts* as offering "computer-based testing to all students" (Michigan and Arkansas) answered "no" to DRC's e-mail questionnaire query, "Does your state offer any statewide computer-delivered testing?"



Source: *Technology Counts 2007*: *A Digital Decade*

The limited implementation of computer-based testing nationwide might be viewed within the overall context of technology in U.S. schools. *Technology Counts* has assigned letter grades to all states based on leadership in three core policy areas: access, use, and capacity. The typical state earned a C-plus (an average score of 77 out of 100 points) with thirty states falling in the C range. South Carolina was given an overall grade of B- (80.1).

Seven of the nine states identified as having at least one statewide assessment available on the computer to all students were given a grade in the A or B range. However, these overall grades do not necessarily correlate with statewide computer-based testing initiatives (e.g., Oregon has been offering its statewide assessments via computer for many years, while its overall technology grade was a D).

Accommodations via Computerized Test Delivery

OVERVIEW

Computerized testing (whether CBT or CAT) has the potential to offer increased flexibility in terms of accommodations for students with disabilities and/or English Language Learners than can be offered via paper/pencil, as long as those accommodations do not change the validity of the test score interpretations. Text-to-speech software, adjustable fonts (style, size, color), single-item per page (screen), spell-checkers, word prediction software, translators, and streaming video are among the tools available via computer—some or all of which can be offered by various test engine vendors, depending on state-defined specifications/allowable accommodations.

When students have been using these "assistive technologies" as part of their daily instructional routine, and the use of these technologies has been documented in a student's Individual Education Plan (IEP) or English Language Learner (ELL) plan, such computer-based accommodations can provide increased equity and access to standardized tests.

A recent study commissioned by the New England Compact Enhanced Assessment Project, "Computer-Based Read-Aloud Tools: An Effective Way to Accommodate Students with Disabilities," found that computer read-aloud accommodations work well for students with disabilities.

- "Students taking the test [high school mathematics] on a computer…did not report major problems with the computer-based test or the read-aloud tools. In addition, the results suggest that the read-aloud method increased students' access to mathematics items. …[S]tudents with special needs performed noticeably better. Students were more willing to listen to an item multiple times on a computer when they might have been reluctant to ask a proctor to re-read an item aloud. The study also found that providing the read-

> aloud accommodation (on computer) closed the gap in performance between native English speakers and students who speak a language other than English at home." (New England Compact, 2005)

However, computer-delivered testing for students with disabilities (or English language learners) may present challenges as well. Many of these challenges are the same as for students without disabilities (e.g., lack of familiarity with answering standardized test questions on a computer screen, using buttons to search for specific items, hand-held calculators vs. calculators displayed on a computer screen, scrolling through long text passages). However, particular care must be taken when accommodating students with disabilities or students who are English language learners to ensure compliance with accommodation guidelines found in both the Individuals with Disabilities Education Act (IDEA) and NCLB legislation.

In testing students with disabilities, it is also important that computer-delivered assessments provide multiple options for selecting responses, such as mouse click, keyboard, touch screen, and assistive devices to access the keyboard (e.g., head wand). See Thompson, Thurlow, & Moore (2003), "Using Computer-based Tests with Students with Disabilities," for further guidance on these issues.

Additionally, as some states have implemented or are investigating "innovative" item types, the challenges of providing access to such items by students with disabilities has proven difficult. For example, since North Carolina's physics end-of-course test includes "innovative item types" that cannot be brailled; an alternate physics assessment had to be created for visually-impaired students. The cost implications to a state should be considered when moving forward in this particular arena.

EXAMPLES OF STATE APPROACHES TO ACCOMMODATIONS VIA COMPUTER-BASED DELIVERY

To illustrate the potential of computer-based accommodations, differing approaches taken by three states are discussed below.

Kentucky

Using various vendors (CTB and Measured Progress), Kentucky has been providing its Core Content Test online (referred to as CATS online) for eligible students for six years. The online test allows the use of Texthelp software (Read and Write Gold, a text-to-speech software program available to all districts in the state for daily instruction), as well as JAWS and ZOOMText software (for students who are blind or have low vision). This computerized assessment is available to students who are English language learners or to students whose IEPs or 504 plans allow for the use of accommodations/assistive technology.

According to the Kentucky State Department of Education website:

- "Many of these students would be unable to participate in the assessment without special accommodations, such as personal readers, scribes (writers), extended time, paraphrasing, use of special technology and equipment, interpreters, etc."
- "It is important to understand that an accommodation does not change the content or difficulty of the test; it reduces the effect of the disability and allows the students to show their knowledge and skills."
- "In the past, approximately 40% of students with disabilities in Kentucky have required personal readers to assist them in taking statewide tests. Over the last few years, Kentucky has greatly reduced the number of students requiring adult assistance by introducing computerized reading supports known as text readers."

NOTE: This online assessment is just one aspect Kentucky's statewide initiative to integrate Universal Design for Learning (UDL) principles to increase the overall achievement of all students, including those with disabilities or with limited English proficiency. This broader initiative has sought to ensure students receive appropriate accommodations in both daily instruction and testing through the procurement and distribution of digital curriculum materials, the creation of an infrastructure of software tools (e.g., text readers), and technical assistance and professional development. (Abell and Lewis, 2004)

As for the assessment-specific impact of this broader initiative, Abell and Lewis found that:

- 89% of students said they believed they did better by taking the test online.
- 82% of students said they could concentrate better.
- 60% of students said they liked that they could work at their own speed.
- Students frequently commented that they felt less embarrassed and intimidated having the computer read questions to them.

Minnesota

This spring the state of Minnesota has implemented a mathematics test for English Language Learners (MTELL) for grades 3–8 and 11 that is delivered via computer. According to the Minnesota Department of Education website, the MTELL is a "plain English" accommodated test for English language learners (ELL) that replaces the MCA-II-Mathematics for AYP accountability. The language in the MTELL is simplified to reduce the reading load for ELL. The test also provides several built-in accommodations that are dependent upon computer delivery, including:

- Simplified language accompanied by graphic supports to allow learners access to unfamiliar vocabulary that is not specific to mathematics.

- Access to audio playback of all written text in the items.

This test became operational in spring 2007 and has not yet been approved by the U.S. Department of Education. However, the Minnesota State Department of Education website indicates the state is confident that the test will pass NCLB peer review.

Massachusetts

A slightly different approach has been taken by the state of Massachusetts. For the Massachusetts Comprehensive Assessment System (MCAS) tests (grades 3–10), the paper test booklet is converted to a "talking test" using Kurzweil 3000 proprietary software. The test is then delivered to eligible students via computer on a CD (not online). Students can track their reading, choose the speed of reading and from several voices, and highlight text. However, students must answer in a standard answer booklet, not on the computer.

NOTE: Of course, there may be specific costs associated with the approaches discussed above, from providing headphones for students using read-aloud accommodations to software licenses at the state, school, or district level. However, some online test engines provide a read-aloud capability that does not require special software on the "user-end"—instead the read-aloud capability is embedded in the testing system. Long-term cost savings may be realized by reducing the personnel needed for providing human readers for individual or small group administrations. This cost savings would be an added benefit to the increased accessibility and performance of students with special needs.

State-Level Standards for Student Keyboarding and Other Technology Skills

As South Carolina considers moving into the CBT/CAT arena, it is appropriate to provide an overview of how states address student standards for technology skills in their curricula.

The recent U.S. Department of Education publication, *State Strategies and Practices for Educational Technology: Volume I—Examining the Enhancing Education through Technology Program* (Bakia, Mitchell, and Yang 2007) (p.2), reports that as of fall 2004:

- Forty-two states had technology standards for students in place
- Of those states, 18 reported "stand-alone" standards
- Sixteen reported embedding technology standards with other academic content standards
- The remaining eight states reported having both stand-alone technology standards and integrated standards.

NOTE: South Carolina has indicated that it does not have stand-alone technology standards for students; rather, International Society for Technology in Education (ITSE) standards are embedded in South Carolina content standards.

Additionally*, Technology Counts* (2007) reports that only four states currently assess students' technology skills: Arizona, Georgia, North Carolina, and Utah.

A sampling of stand-alone state technology standards for students reveals some variation among the states.

Texas, for example, does not offer online testing for students at the elementary grades. However, its technology standards for K–2 require, in part, the following:

a. use a variety of input devices such as mouse, keyboard, disk drive, modem, voice/sound recorder, scanner, digital video, CD-ROM, or touch screen;

b. use proper keyboarding techniques such as correct hand and body positions and smooth and rhythmic keystroke patterns as grade-level appropriate;

c. demonstrate touch keyboarding techniques for operating the alphabetic, numeric, punctuation, and symbol keys as grade-level appropriate;

d. produce documents at the keyboard, proofread, and correct errors

e. use language skills including capitalization, punctuation, spelling, word division, and use of numbers and symbols as grade-level appropriate.

Source: *Texas Essential Knowledge and Skills* (1998) Texas Education Agency

Virginia, which does offer computer-based testing starting at Grade 3 (without writing or constructed response), requires similar skills of its K–2 students:

C/T K-2.2 The student will demonstrate proficiency in the use of technology.

- Demonstrate the use of mouse, keyboard, printer, multimedia devices, and earphones.
- Use multimedia resources such as interactive books and software with graphical interfaces.

C/T K-2.7 The student will use a variety of media and formats to communicate information and ideas effectively to multiple audiences.

- Identify the best tool to communicate information.
- Use technology tools for individual writing, communication, and publishing activities.
- Demonstrate the ability to create, save, retrieve, and print documents.

Source: *Standards of Learning* (2005), Virginia Department of Education

North Carolina, which does not offer computer-based testing below grade 8, breaks out its technology standards by grade, rather than grade cluster. The state's technology standards for grade 3 include:

- 1.09 Identify and use formatting terms/concepts (e.g., font size/style, line spacing, margins, italic). (4)
- 2.03 Use prepared databases to search/filter and sort alphabetically/numerically in ascending/descending order. (2)
- 2.07 Demonstrate correct finger placement for home row keys. (4)
- 2.08 Use menu/tool bar functions (e.g., font size/style, line spacing) to format and change the appearance of word processing documents as a class/group. (4)
- 3.02 Enter/edit data in a prepared spreadsheet to perform calculations and determine which graph best represents the data as a class/group. (3)
- 3.03 Use word processing as a tool to write, edit, and publish sentences, paragraphs, and stories. (4)

Source: *Standard Course of Study*, Computer/Technology Skills (n.d.), North Carolina Department of Public Instruction.

Other states break out their technology standards by K–4 vs. K–2 or by grade. For example, **New Jersey** requires, in part, that, by the end of Grade 4, students will:

- Use basic features of an operating system (e.g., accessing programs, identifying and selecting a printer, finding help).
- Input and access text and data, using appropriate keyboarding techniques or other input devices.

- Produce a simple finished document using word processing software.
- Produce and interpret a simple graph or chart by entering and editing data on a prepared spreadsheet template.
- Create and present a multimedia presentation using appropriate software.
- Create and maintain files and folders.

Source: *Core Curriculum Content Standards*, Technological Literacy (2004), New Jersey Department of Education.

The extent to which any or all of the standards listed above are implemented at a school or classroom level is unknown. However, state-mandated standards for technology skills, in conjunction with practice tests and tutorials, may help to ensure student success with any CBT/CAT testing initiative.

TEACHER TECHNOLOGY SKILLS STANDARDS

A related issue is state-level teacher technology standards. According to the same USDOE report cited above, as of fiscal year (FY) 2003 "50 percent of states were assessing or planning to assess teacher technology skills or were relying on districts to do so" (Bakia, Mitchell, and Yang 2007, p. 29). South Carolina reports that it has minimum technology standards for teachers, but does not currently formally assess those technology skills, nor does it report plans to do so (p. 30). Twenty-seven states report offering online professional development for teachers to help them integrate technology into core subject areas (p. 31). A state-by-state breakout is not provided.

These findings also suggest the importance of professional development and training for teachers who may be involved in CBT/CAT.

Hardware, Software, and Security Issues

NOTE: To avoid redundancy in this report, the issues of hardware and software (system requirements) of various e-testing vendors are addressed as part of Component 6; security considerations are addressed as part of Component 10.

Reporting

Please see questionnaire responses below for state-specific reporting of CBT/CAT tests. A fuller discussion of reporting in general, testing windows, report turnaround times, item development costs for releasing items, and formative assessments is included in Component 11.

State Responses to E-mail Questionnaire Regarding CBT/CAT Testing

Alaska

Does your state offer any statewide computer-delivered testing?

- Yes—Formative Assessment, the Alaska Computerized Formative Assessment (ACFA)

What grade and subject-area tests are offered via computer?

- Grades 3–10 Subjects: Reading and Mathematics

Is testing via computer voluntary or mandatory?

- Voluntary

If voluntary, about what percent of testing takes place via computer?

- Grades 3–10 Subjects: Reading and Mathematics

Approximately how many computer-delivered tests were administered in your state last year (i.e., number of tests delivered via computer)?

- We have had 3,141 test administrations created, and 18,870 students that have taken these tests.

Are the computer-delivered tests "fixed" or "scrambled" forms, OR computer-adaptive?

- Fixed forms

Do the computer-delivered tests include short-constructed responses?

- No

Do the computer-delivered tests include extended-writing responses?

- No

How quickly are computer-delivered test scores made available to the student, school, and district (i.e., what is the reporting turnaround time for these different groups)?

- Once students have completed testing, preliminary results are available immediately. Preliminary results show the correct/incorrect responses. Advanced detailed results will be available for viewing no later than the morning after the assessment was given.
- Reports currently available:
  1. Student Score Listing – Alphabetical
  2. Student Score Listing - Highest to Lowest
  3. Student Problem Table with Modified Caution Index. (This report shows how typical student's correct item responses are in relation to the rest of the students in the same building on the formative test. The modified caution index may range from zero to 99.)
  4. Item Performance Summary
  5. Group Report
- Reports are not available at a building or a district level as this is not the intent of formative assessments.

Which vendor(s) are providing your test engine(s) for your computer-delivered testing?

- Computerized Assessments and Learning, LLC (CAL), as subcontractor to Data Recognition Corporation.

Is there anything else you would like to share from your state's experience in implementing computer-delivered testing that might inform the State of South Carolina's Computer-Based/Computer-Adaptive Feasibility Study?

- Districts and schools within districts have varying levels of connectivity; reliable connections are unpredictable
- Varying numbers of computers in districts and in schools within districts; scheduling time to take the tests can be challenging
- When 25–50 users were on the system at the same time, the response from the site became very slow
- Immediate delivery of scores from tests did not necessarily result in teachers using those scores to adjust instruction
- Challenging to make any changes to a pre-existing system (i.e., a system that has been adapted to serve our state)
- I would also share that our state board is very interested in having us move the operational part of our tests to online, but we have not gone there as we are looking to see how the formative assessment program works online.

CONNECTICUT

Does your state offer any statewide computer-delivered testing?

- Only limited, at the moment, to one part of the test. The Skills Checklist is administered in an online environment. The "test" is completed by an adult, not the student, and turned in online. Less than 1% of the students are assessed using the Skills Checklist. Other than that, there is no "Statewide" testing online at the moment.

What grade and subject-area tests are offered via computer?

- 3, 4, 5, 6, 7, 8, and 10. Skills Checklists are in mathematics, language arts, and science. The science is being piloted.

Is testing via computer voluntary or mandatory?

- Mandatory for this small subset of test documents.

Approximately how many computer-delivered tests were administered in your state last year (i.e., number of tests delivered via computer)?

- Zero. This is our first year requiring the computer interface.

Are the computer-delivered tests "fixed" or "scrambled" forms, OR computer-adaptive?

- Fixed

Do the computer-delivered tests include short-constructed responses?

- No

Do the computer-delivered tests include extended-writing responses?

- No

How quickly are computer-delivered test scores made available to the student, school, and district (i.e., what is the reporting turnaround time for these different groups)?

- They are folded into the results for the non-computer delivered tests, so it takes fully 3 months after testing to receive results. Though these data are known right away, they are reported later.

Which vendor(s) are providing your test engine(s) for your computer-delivered testing?

- Measurement Incorporated in Durham, N.C.

Is there anything else you would like to share from your state's experience in implementing computer-delivered testing that might inform the State of South Carolina's Computer-Based/Computer-Adaptive Feasibility Study?

- We are otherwise engaged in piloting on a small scale some portions of our grade 10 test.

If your state does not currently offer any statewide computer-delivered testing, is your state planning (or has conducted) any pilots or feasibility studies of computer-delivered testing?

- We have plans to convert our test into a computer ready version in the future. Piloting will give us part of the answer to the feasibility question. There are certainly questions of access and load that are not yet answered in Connecticut.

Florida

Does your state offer any statewide computer-delivered testing?

- Yes

What grade and subject-area tests are offered via computer?

- Re-takers of the Grade 10 FCAT Reading and Mathematics

  (Grades 10, 11, 12, 13, AD)

Is testing via computer voluntary or mandatory?

- Voluntary

  If voluntary, about what percent of testing takes place via computer?

- The first time, fall 2006, less than 1% tested on the computer; a higher percentage is testing this spring 2007.

Approximately how many computer-delivered tests were administered in your state last year (i.e., number of tests delivered via computer)?

- 240

Are the computer-delivered tests "fixed" or "scrambled" forms, OR computer-adaptive?

- Scrambled

Do the computer-delivered tests include <u>short-constructed</u> responses?

- No

Do the computer-delivered tests include <u>extended-writing</u> responses?

- No

How quickly are computer-delivered test scores made available to the student, school, and district (i.e., what is the reporting turnaround time for these different groups)?

- Approximately 6 weeks for all data for the Retake administrations

Which vendor(s) are providing your test engine(s) for your computer-delivered testing?

- CTB/McGraw-Hill

Is there anything else you would like to share from your state's experience in implementing computer-delivered testing that might inform the State of South Carolina's Computer-Based/Computer-Adaptive Feasibility Study?

- [No Response—See comments provided during Expert Panel Meeting]

GEORGIA

Does your state offer any statewide computer-delivered testing?

- End of Course Tests are offered both online and paper/pencil. School systems make a local decision as to whether they want to administer some or all of their EOCT's online.

What grade and subject-area tests are offered via computer?

- All eight EOCT's are offered online for each administration. These tests typically include students in grades 9–12, and may include small numbers of students in grades 7–8 in some systems. EOCT's exist in Lang Arts, Science Social Studies, and Math.

Is testing via computer voluntary or mandatory?

- Voluntary

If voluntary, about what percent of testing takes place via computer?

- Approximately 15%–20%

Approximately how many computer-delivered tests were administered in your state last year (i.e., number of tests delivered via computer)?

- Spring 2006 and Winter 2006 combined included 131,910 online EOCT's

Are the computer-delivered tests "fixed" or "scrambled" forms, OR computer-adaptive?

- Fixed

Do the computer-delivered tests include short-constructed responses?

- No

Do the computer-delivered tests include extended-writing responses?

- No

How quickly are computer-delivered test scores made available to the student, school, and district (i.e., what is the reporting turnaround time for these different groups)?

- ***Student results:*** 5 days from date of test completion
- ***School results:*** Summary results (online and paper/pencil combined) are available at the end of the state testing window
- ***District results:*** Same as above
- ***Statewide results:*** Same as above

Which vendor(s) are providing your test engine(s) for your computer-delivered testing?

- Pearson Educational Measurement

IDAHO

Does your state offer any statewide computer-delivered testing?

- Yes

What grade and subject-area tests are offered via computer?

- Grades 2–10 in Reading, Math and Language Usage; Grades 5, 7, 10 in Science

Is testing via computer voluntary or mandatory?

- Mandatory

Approximately how many computer-delivered tests were administered in your state last year (i.e., number of tests delivered via computer)?

- About 20,000 students per grade (600,000 total)

Are the computer-delivered tests "fixed" or "scrambled" forms, OR computer-adaptive?

- Idaho is a combination for the spring administration; there are fixed sections, scrambled sections, and adaptive pieces. For the fall and winter administrations, they will be completely adaptive.

Do the computer-delivered tests include short-constructed responses?

- No

Do the computer-delivered tests include extended-writing responses?

- No

How quickly are computer-delivered test scores made available to the student, school, and district (i.e., what is the reporting turnaround time for these different groups)?

- ***Student results:*** after this spring admin—immediate student response.
- ***School results:*** after this spring admin—once all students finish testing
- ***District results:*** after this spring admin—once all students finish testing
- ***Statewide results:*** after this spring admin—early June

Which vendor(s) are providing your test engine(s) for your computer-delivered testing?

- Computerized Assessments and Learning, LLC (CAL), as subcontractor to Data Recognition Corporation.

Is there anything else you would like to share from your state's experience in implementing computer-delivered testing that might inform the State of South Carolina's Computer-Based/Computer-Adaptive Feasibility Study?

- Idaho has been happy with our computer-delivered tests. Change is always hard, and some school districts were very apprehensive when we started the computer-delivered tests. At this point, however, there would not be one school district who would wish to go back to paper and pencil. They are easier to administer and results come back in a timely manner. Those results can then be used for instruction purposes.
  There will be some up-front costs associated with converting to computer-delivered testing. However, once the up-front costs have been paid, the cost of computer-delivered testing is equal to paper-pencil testing. Recommend that South Carolina move in the direction of computer-delivered tests.

INDIANA

Does your state offer any statewide computer-delivered testing?

- Yes

What grade and subject-area tests are offered via computer?

- Algebra I, Algebra II, English 11, Biology I

Is testing via computer voluntary or mandatory?

- During the winter administrations, only CBT is available. During the spring administration, CBT or PP is available—schools may select by content area.

If voluntary, about what percent of testing takes place via computer?

- Last year's numbers by subject: Algebra I: 59,551 or 81%; English 11: 49,083 or 87%

Approximately how many computer-delivered tests were administered in your state last year (i.e., number of tests delivered via computer)?

- 108,634 operational assessments in Algebra I and English 11 combined; 20,104 Biology I pilot assessments; and 11,255 Algebra II pilot assessments

Are the computer-delivered tests "fixed" or "scrambled" forms, OR computer-adaptive?

- Fixed

Do the computer-delivered tests include <u>short-constructed</u> responses?

- Yes

If yes, in what grades and subjects are these offered?

- Algebra I, Algebra II, Biology I, English 11

How are these responses scored?

- By hand. Online tests are delivered electronically [for] human scoring. PP tests are scanned and sent electronically to same group of human scorers.

Do the computer-delivered tests include extended-writing responses?

- Yes

If yes, in what grades and subjects are these offered?

- English 11 essay only

How are these responses scored?

- Essays are sent for human scoring only if pushed out by AI scoring engine

How quickly are computer-delivered test scores made available to the student, school, and district (i.e., what is the reporting turnaround time for these different groups)?

- 24 hrs from close of student test[ing window]; after 10 days of testing

Which vendor(s) are providing your test engine(s) for your computer-delivered testing?

- Questar Assessment Inc, formerly known as Achievement Data Incorporated

Is there anything else you would like to share from your state's experience in implementing computer-delivered testing that might inform the State of South Carolina's Computer-Based/Computer-Adaptive Feasibility Study?

- Special challenges include finding a vendor that is a one-stop shop for delivering, scoring, reporting, and psychometric services for CBT and PP assessments (simultaneously).

KANSAS

Does your state offer any statewide computer-delivered testing?

- Yes

What grade and subject-area tests are offered via computer?

- Reading 3–8, high school grade 11
- Mathematics 3–8, plus grade 11
- Social Studies 6, 8, and 10

Is testing via computer voluntary or mandatory?

- Voluntary

If voluntary, about what percent of testing takes place via computer?

- Approximately 75%

Approximately how many computer-delivered tests were administered in your state last year (i.e., number of tests delivered via computer)?

- Approximately 335,000

Are the computer-delivered tests "fixed" or "scrambled" forms, OR computer-adaptive?

- Five active forms with each form having different items—all fixed forms.

Do the computer-delivered tests include short-constructed responses?

- No

Do the computer-delivered tests include extended-writing responses?

- No

How quickly are computer-delivered test scores made available to the student, school, and district (i.e., what is the reporting turnaround time for these different groups)?

- ***Student results:*** Either immediate or at the latest, by the next morning following completion of the last test part, depending on testing load.
- ***School results:*** Either immediate or at the latest, by the next morning following completion of the last test part, depending on testing load.
- ***District results:*** Not made available until all testing, both computer and paper/pencil, are completed. Unofficial reporting in mid- to late May.
- ***Statewide results:*** Not available until August.

Which vendor(s) are providing your test engine(s) for your computer-delivered testing?

- Center for Educational Testing and Evaluation (CETE), University of Kansas.

Is there anything else you would like to share from your state's experience in implementing computer-delivered testing that might inform the State of South Carolina's Computer-Based/Computer-Adaptive Feasibility Study?

- Began small, one grade and content area…then some grades and two areas; before going [to] all grades and content areas. And use research to build the case for equity, fairness, comparability, and soundness.

Kentucky

Does your state offer any statewide computer-delivered testing?

- Yes. Special Populations students whose IEP or 504 Plan whose daily instruction allows for the use of computer/assistive technology can take the statewide summative assessment (CATS) as a computer delivered assessment. We will conduct another pilot with general population students in spring 2008.

What grade and subject-area tests are offered via computer?

- Grades 3 – 8, 10, and 11. Content areas are reading, mathematics, science, social studies, arts & humanities, practical living vocational studies, and on-demand writing. Reading and mathematics are grade specific for NCLB; other content areas are administered once in El, MS, and HS.

Is testing via computer voluntary or mandatory?

- Voluntary and only those students identified in their IEP or 504 plan.

Approximately how many computer-delivered tests were administered in your state last year (i.e., number of tests delivered via computer)?

- Approximately 2,000 students.

Are the computer-delivered tests "fixed" or "scrambled" forms, OR computer-adaptive?

- Fixed

Do the computer-delivered tests include short-constructed responses?

- YES – Students respond to prompts and have the equivalent of a full page to answer. (On-demand writing is 4 pages.) Number of Open Response questions varies by content area and ranges from 2 to 7.

If yes, in what grades and subjects are these offered?

- All content areas at grade's content area are assessed. [according to KY website, same as for the Kentucky Core Content Test (KCCT)]

How are these responses scored?

- Direct delivery via online and then scored by human readers working for the Testing Contractor.

Do the computer-delivered tests include extended-writing responses?

- See above

How are these responses scored?

- See above

How quickly are computer-delivered test scores made available to the student, school, and district (i.e., what is the reporting turnaround time for these different groups)?

- Results are reported with paper and paper results with a goal of August 1. Kentucky's goal is to test all students online by 2012. At [this] time our intention is [to] release multiple choice findings within a couple of weeks of the test administration. We also plan to release not scored student common open response (vs. matrix) images back to the schools.

Which vendor(s) are providing your test engine(s) for your computer-delivered testing?

- 2007 – CTB and eCollege
- 2008 – Measured Progress

Is there anything else you would like to share from your state's experience in implementing computer-delivered testing that might inform the State of South Carolina's Computer-Based/Computer-Adaptive Feasibility Study?

- Kentucky plans to expand testing students online with Formative and Benchmark Assessments. Development of end-of-course assessments at the high school level is underway, and these will be offered in a combination of Online and paper and pencil.

Maryland

Does your state offer any statewide computer-delivered testing?

- Yes

What grade and subject-area tests are offered via computer?

- Science tests in grades 5 and 8 are offered online. This test is new and is being administered for the first time between April 18 and May 8, 2007.

Is testing via computer voluntary or mandatory?

- The science tests are census tests and are mandatory; however, the tests are not mandated to be taken via computer and may be taken via paper.

If voluntary, about what percent of testing takes place via computer?

- We expect that approximately 40% of the students testing in science in grades 5 and 8 in 2007 will take the tests online.

Approximately how many computer-delivered tests were administered in your state last year (i.e., number of tests delivered via computer)?

- None

Are the computer-delivered tests "fixed" or "scrambled" forms, OR computer-adaptive?

- The science tests are 10 fixed forms per grade.

Do the computer-delivered tests include short-constructed responses?

- Yes

If yes, in what grades and subjects are these offered?

- The science tests in grades 5 and 8 both include Brief Constructed Response items.

How are these responses scored?

- For operational purposes, all scoring is done by human readers. [However,] our state is conducting a pilot study using AI [Artificial Intelligence] scoring as the second read. Maryland will evaluate the results of this study to determine whether to proceed to any use of AI scoring for operational purposes.

Do the computer-delivered tests include extended-writing responses?

- No

How quickly are computer-delivered test scores made available to the student, school, and district (i.e., what is the reporting turnaround time for these different groups)?

- As 2007 is the first administration of the science test, we have many unique tasks to complete in the first year, including standard setting, which will impact the timing of results. We do not expect to report any results from the April/May 2007 testing until at least October 2007. In future years, we would hope to return results in the summer following test administration.
- ***Student results:*** N/A for 2007 — see above
- ***School results:*** N/A for 2007 — see above
- ***District results:*** N/A for 2007 — see above
- ***Statewide results:*** N/A for 2007 — see above

Which vendor(s) are providing your test engine(s) for your computer-delivered testing?

- We use Pearson Educational Measurement to provide the online testing engine.

Is there anything else you would like to share from your state's experience in implementing computer-delivered testing that might inform the State of South Carolina's Computer-Based/Computer-Adaptive Feasibility Study?

- No

Minnesota

Does your state offer any statewide computer-delivered testing?

- Yes

What grade and subject-area tests are offered via computer?

- Mathematics for English language learners — grades 3–8 and 11.
- Minnesota Comprehensive Assessment (MCA-II) in Science — grades 5, 8 and once in high school (after completion of life sciences curriculum)

Is testing via computer voluntary or mandatory?

- MTELL is voluntary; students will be able to take it instead of the regular Mathematics test. All students in grades 3–8 and 11 must take either MTELL or the MCA-II-Mathematics.
- MCA-II-Science will be mandatory in the spring of 2008. Participation in field testing spring 2006 and spring 2007 has been voluntary.

If voluntary, about what percent of testing takes place via computer?

- So far only field testing for both of these tests has been conducted. MTELL will first become operational in spring 2007.

Approximately how many computer-delivered tests were administered in your state last year (i.e., number of tests delivered via computer)?

- Only field testing in a limited # of districts so far.

Are the computer-delivered tests "fixed" or "scrambled" forms, OR computer-adaptive?

- Fixed within forms

Do the computer-delivered tests include short-constructed responses?

- Yes

If yes, in what grades and subjects are these offered?

- Math and science; for all grades covered by these tests

How are these responses scored?

- A combination (by computer, by hand). In math, a student fills in a grid with numbers and these are scored by computers. With Science, the scoring method depends on the nature of the response and whether trained (human) raters are needed.

Do the computer-delivered tests include extended-writing responses?

- Yes

If yes, in what grades and subjects are these offered?

- Science tests

How are these responses scored?

- "By hand" — by trained (human) raters.

How quickly are computer-delivered test scores made available to the student, school, and district (i.e., what is the reporting turnaround time for these different groups)?

- Turnaround time will be 2 to 3 months for all results.

Which vendor(s) are providing your test engine(s) for your computer-delivered testing?

- Pearson Educational Measurement

Is there anything else you would like to share from your state's experience in implementing computer-delivered testing that might inform the State of South Carolina's Computer-Based/Computer-Adaptive Feasibility Study?

- One major limiting factor for moving to online testing has been the availability of computers in the schools, since they are often already booked solid.
- Once a test goes online, there is not a comparable paper and pencil test that can be used, due to cost of developing 2 tests and comparability issues.
- The plan to move tests to computer delivery has been scaled back, based on district input that they do not have the resources. Important to assess resources and get input from district staff early in the process.
- Using computer-adaptive testing for statewide accountability testing is problematic, although there has been some push for this.

If your state does not currently offer any statewide computer-delivered testing, is your state planning (or has conducted) any pilots or feasibility studies of computer-delivered testing? (If yes, please explain.)

- Small scale field testing has been conducted. MTELL is scheduled to go operational this spring, science in spring 2008.

MISSISSIPPI

Does your state offer any statewide computer-delivered testing?

- Yes

What grade and subject-area tests are offered via computer?

- End of course tests in Algebra I, Biology I, English II, and U.S. History from 1877

Is testing via computer voluntary or mandatory?

- Re-tests are all online. First time testers take paper and pencil.

Approximately how many computer-delivered tests were administered in your state last year (i.e., number of tests delivered via computer)?

- Algebra I: 2,199
- Biology: 1,727
- English II Multiple Choice: 6,329
- U.S. History: 831
- Algebra I: 1,357
- Biology: 1,111
- English II Multiple Choice: 4,230
- US History: 503

Are the computer-delivered tests "fixed" or "scrambled" forms, OR computer-adaptive?

- Fixed

Do the computer-delivered tests include short-constructed responses?

- No

Do the computer-delivered tests include extended-writing responses?

- No

How quickly are computer-delivered test scores made available to the student, school, and district (i.e., what is the reporting turnaround time for these different groups)?

- ***Student results:*** 3 weeks for Pass/Fail Rosters – Student score reports 5 to 6 weeks. They only receive individual results for online re-tests. Re-testers are not part of the state accountability system, only first time testers are included.
- ***School results:*** NA
- ***District results:*** NA
- ***Statewide results:*** NA

Which vendor(s) are providing your test engine(s) for your computer-delivered testing?

- Harcourt Assessment, Inc.

Is there anything else you would like to share from your state's experience in implementing computer-delivered testing that might inform the State of South Carolina's Computer-Based/Computer-Adaptive Feasibility Study?

- We have been testing our re-testers online for four years and have been quite successful. With each administration, different problems have occurred, but the administrations have been successful. The key is to have a vendor that is willing to work with you and talk through the problems and for solutions to be implemented as quickly as possible.
- The majority of students prefer online testing.

North Carolina

Does your state offer any statewide computer-delivered testing?

- Yes

What grade and subject-area tests are offered via computer?

- Online Test of Computer Skills, grade 8; 8 High School End-of-Course tests (algebra 1, geometry, algebra 2; English 1; biology, physics; U.S. history, civics & economics); online data collection for an alternate assessment (1%) and an alternative assessment (primarily LEP but also certain categories of disability, i.e., newly blinded). The teachers enter their data; the kids do not use the computer; also doing a special study for feasibility of writing test online.

Is testing via computer voluntary or mandatory?

- Computer Skills, mandatory (graduation test); there is an alternative, but fewer than 25 students are on it; EOCs optional except Physics; Physics is a field test and is only online. Alternate and alternative are mandatory for teachers to enter their data; Special studies mandatory for sampled schools.

If voluntary, about what percent of testing takes place via computer?

- This spring is the first time we've offered it on a large scale. We asked for a voluntary sample to double-test in paper/pencil AND online environment, and we have about 10,000 volunteers.

Approximately how many computer-delivered tests were administered in your state last year (i.e., number of tests delivered via computer)?

- Computer skills, 130,000

Are the computer-delivered tests "fixed" or "scrambled" forms, OR computer-adaptive?

- Fixed

Do the computer-delivered tests include <u>short-constructed</u> responses?

- Kind of ... the computer skills [test] has performance items (like do a task within an application environment); the physics has "simulations" where the student can change variables to meet a set of criteria or respond to a question.

If yes, in what grades and subjects are these offered?

- [see above]

How are these responses scored?

- Pattern-matching or captured keystrokes; all scored dichotomously

Do the computer-delivered tests include <u>extended-writing</u> responses?

- Only the special studies for writing

How are these responses scored?

- I think part of the study was to compare hand to AI scoring as well as just delivery

How quickly are computer-delivered test scores made available to the student, school, and district (i.e., what is the reporting turnaround time for these different groups)?

- Student results—after initial embargo while data are verified, generally next day. State-provided software can produce multiple levels and types of reports, including classroom, school, and district (LEA). All data are brought in to the state at the end of June; final reporting results are generally disseminated in October.

Which vendor(s) are providing your test engine(s) for your computer-delivered testing?

- Do not use a vendor—we have a partnership with a state university, and part of that includes technological support.

Is there anything else you would like to share from your state's experience in implementing computer-delivered testing that might inform the State of South Carolina's Computer-Based/Computer-Adaptive Feasibility Study?

- If you do CAT, are you going item-wise or with "chunks" of items? We tried chunks (boil/freeze CAT) and saw no increase in reliability or decrease in SEM as would have been theoretically expected.

If your state does not currently offer any statewide computer-delivered testing, is your state planning (or has it conducted) any pilots or feasibility studies of computer-delivered testing? (If yes, please explain.)

- Did feasibility on delivery mode, using teachers. Did additional feasibility of interface, using students and teachers. Currently doing comparability studies for the EOCs.

Oklahoma

Does your state offer any statewide computer-delivered testing?

- Yes

What grade and subject-area tests are offered via computer?

- For the grade 3–8 program, grade 7 geography has been online since 2004. Grade 8 math and reading will be administered online beginning in 2008.
- For the End of Instruction (EOI) program, all 7 assessments will be offered online by 2008-9. The seven assessments are Algebra I, Algebra II, English II, English III, Geometry, US History, and Biology I.
- For spring 2007, only Algebra II, English III, and Geometry are being field tested online.
- Biology I and U.S. History End of Instruction assessments were offered online in the spring 2006 and winter 2006.

Is testing via computer voluntary or mandatory?

- For the grade 3–8 program, online testing is mandatory, unless the paper version is needed as an accommodation for IEP students.

Approximately how many computer-delivered tests were administered in your state last year (i.e., number of tests delivered via computer)?

- For the 3–8 program, approximately 45,000 students were administered the online assessment.
- For the EOI program, less than 50% of the eligible students took the assessments online. Approximately 45,000 students are eligible for the tests.

Are the computer-delivered tests "fixed" or "scrambled" forms, OR computer-adaptive?

- Scrambled forms are used.

Do the computer-delivered tests include short-constructed responses?

- No

Do the computer-delivered tests include extended-writing responses?

- No

How quickly are computer-delivered test scores made available to the student, school, and district (i.e., what is the reporting turnaround time for these different groups)?

- ***Student results:*** Raw scores are made available immediately after testing.
- ***School results:*** Online Preliminary Reports are available two weeks after the close of the testing window.
- Hardcopy preliminary reports are available by June 1.
- Hardcopy final reports are available by July 1.

Which vendor(s) are providing your test engine(s) for your computer-delivered testing?

- For the grade 3-8 program, the vendor is DRC. The online engine is provided by Internet Testing Systems (ITS), as subcontractor to ETS, as subcontractor to DRC.
- For the End of Instruction Program, Pearson is the current vendor and provided the online testing system. In past years, CTB was the vendor for this program.

Is there anything else you would like to share from your state's experience in implementing computer-delivered testing that might inform the State of South Carolina's Computer-Based/Computer-Adaptive Feasibility Study?

- The importance of training for district personnel cannot be over-estimated. It is most beneficial if the vendor can provide a "system check" that will allow districts to run an application to evaluate each computer to determine if minimum system requirements are met. Districts will need guidance from the state regarding scheduling of testing to get their entire student population testing within the testing window.
- The state should put policies in place that deal with technical difficulties that may arise during testing due to severe weather or other disaster. Consider having printed equivalent forms available.
- Make sure your vendor has appropriate technical support staff available to assist districts before and during testing.

Oregon

Please note: responses were submitted prior to Oregon's reversion to paper/pencil testing for Spring 2007.

Does your state offer any statewide computer-delivered testing?

- Yes

What grade and subject-area tests are offered via computer?

- Math, Science, reading, Grades 3–12
- English Proficiency grades K–12

Is testing via computer voluntary or mandatory?

- Mandatory, approximately 1–2% of students still need paper and pencil assessments

Approximately how many computer-delivered tests were administered in your state last year (i.e., number of tests delivered via computer)?

- 1.2 million

Are the computer-delivered tests "fixed" or "scrambled" forms, OR computer-adaptive?

- English Proficiency Tests are fixed. Academic tests are adaptive (constrained by grade level and score reporting categories).

Do the computer-delivered tests include short-constructed responses?

- Yes

If yes, in what grades and subjects are these offered?

- English Proficiency Assessments (K–12)

How are these responses scored?

- Some short spoken sentences are scored by AI, all other responses scored manually.

Do the computer-delivered tests include extended-writing responses?

- No

How quickly are computer-delivered test scores made available to the student, school, and district (i.e., what is the reporting turnaround time for these different groups)?

- ***Student results:*** Immediately
- ***Statewide results:*** August of each year for systems level reports.

Which vendor(s) are providing your test engine(s) for your computer-delivered testing?

- Vantage -Academic
- Language Learning Solutions -English proficiency

Is there anything else you would like to share from your state's experience in implementing computer-delivered testing that might inform the State of South Carolina's Computer-Based/Computer-Adaptive Feasibility Study?

- [No Response]

SOUTH CAROLINA

Does your state offer any statewide computer-delivered testing?

- Yes

What grade and subject-area tests are offered via computer?

- The End-of-Course tests are offered online. End-of-Course tests are administered to students in middle and high school who take the following courses for high school credit: Algebra 1/Mathematics for the Technologies 2, English 1, Physical Science, US History and Constitution.

Is testing via computer voluntary or mandatory?

- Voluntary at the student level.

If voluntary, about what percent of testing takes place via computer?

- 27% took the test via computer in fall 2006

Approximately how many computer-delivered tests were administered in your state last year (i.e., number of tests delivered via computer)?

- 32,750 tests were administered online in 2006 (spring 2006, summer 2006, and December 2006/January 2007).

Are the computer-delivered tests "fixed" or "scrambled" forms, OR computer-adaptive?

- Multiple fixed forms

Do the computer-delivered tests include short-constructed responses?

- No, they are multiple choice only

How quickly are computer-delivered test scores made available to the student, school, and district (i.e., what is the reporting turnaround time for these different groups)?

- ***Student results:*** reported online within 36 hours after the student completes the test
- ***School results:*** school summaries - 4 to 5 weeks after the state testing window closes (these are provided on paper and CD)
- ***District results:*** district summaries - 4 to 5 weeks after the state testing window closes (these are provided on paper and CD)
- ***Statewide results:*** state summaries - approximately 4 weeks after the state window closes (these are provided on paper and CD)

Which vendor(s) are providing your test engine(s) for your computer-delivered testing?

- Pearson Educational Measurement

Is there anything else you would like to share from your state's experience in implementing computer-delivered testing that might inform the State of South Carolina's Computer-Based/Computer-Adaptive Feasibility Study?

- [No Response]

South Dakota

Does your state offer any statewide computer-delivered testing?

- Yes, we have two computer based tests—DACS (Dakota Assessment of Content Standards) and Achievement Series. [NOTE: DACS and Achievement Series are both formative in nature and neither are used for NCLB purposes. The Dakota STEP is our NCLB assessment.]

What grade and subject-area tests are offered via computer?

- DACS (Dakota Assessment of Content Standards) Math, Reading Science and Lang Arts Grades 2–12
- Achievement Series-Reading and Math K–11

Is testing via computer voluntary or mandatory?

- Voluntary

If voluntary, about what percent of testing takes place via computer?

- In Performance Series 100% of the original test is taken on the computer.

Approximately how many computer-delivered tests were administered in your state last year?

- Performance Series = 64,596

Are the computer-delivered tests "fixed" or "scrambled" forms, OR computer-adaptive?

- DACS is computer-adaptive and AS is fixed form.

Computer adaptive (i.e., the test items are chosen and presented to a student based on that student's response to previous item)

- This one for Performance Series

Do the computer-delivered tests include short-constructed responses?

- [No response]

Do the computer-delivered tests include extended-writing responses?

- [No response]

How quickly are computer-delivered test scores made available to the student, school, and district (i.e., what is the reporting turnaround time for these different groups)?

- For both formative assessments, DACS and Achievement Series, results are:
- ***Student results:*** as soon as the student finishes a test
- ***School results:*** as soon as student takes test
- ***District results:*** as soon as student takes test
- ***Statewide results:*** as soon as student takes test

Which vendor(s) are providing your test engine(s) for your computer-delivered testing?

- Performance Series (Scantron)

TEXAS

Does your state offer any statewide computer-delivered testing?

- Yes, we continue to expand online testing each year and plan to offer at least one version of all our statewide assessments online within the next several years. We also have several assessments that are currently offered exclusively online.
- For more information see the letter on our website http://www.tea.state.tx.us/student.assessment/resources/letters/2007/061201_Online_testing_plans_with_letterhead.pdf
- [Gleaned from the letter referenced above: Fall 2006, Exit Level Retest; Spring 2007, Grades 7, 8, 9, 10, and Exit Level Retest—all grades, all tested subjects; Spring 2008, adding Grades 5 and 6; Spring 2009, adding Grades 3 and 4. Spring 2009, adding Reading Proficiency Test in English II, Grades 2-12—to be available only online. Phasing-in six End-of-Course tests, not state-mandated participation, but only offered online.]

What grade and subject-area tests are offered via computer?

- Please see the letter on our website for complete information about current and long-range plans: http://www.tea.state.tx.us/student.assessment/resources/letters/2007/061201_Online_testing_plans_with_letterhead.pdf
- [See information gleaned from referenced letter under previous question.]

Is testing via computer voluntary or mandatory?

- Currently, only certain online field tests require mandatory participation.

If voluntary, about what percent of testing takes place via computer?

- About 10% participate in our voluntary administrations.

Approximately how many computer-delivered tests were administered in your state last year (i.e., number of tests delivered via computer)?

- About 50,000 total online tests

Are the computer-delivered tests "fixed" or "scrambled" forms, OR computer-adaptive?

- Our tests are fixed forms, not scrambled.

Do the computer-delivered tests include short-constructed responses?

- Yes — 9 grade Reading, 10th and 11th Grade English Language Arts

How are these responses scored?

- The responses are scored by human readers after being typed in by the student in TestNav [Pearson test engine]. The responses are then electronically transferred to the [Pearson] ePen system for scoring.

Do the computer-delivered tests include extended-writing responses?

- Yes — 10th and 11th Grade English Language Arts tests

How are these responses scored?

- By hand

How quickly are computer-delivered test scores made available to the student, school, and district (i.e., what is the reporting turnaround time for these different groups)?

- At this time, we follow the same schedule as the paper assessments in providing scoring data of district, school, and student results - 2 weeks after the administration. Statewide results are provided annually.

Which vendor(s) are providing your test engine(s) for your computer-delivered testing?

- Pearson Educational Measurement

UTAH

Does your state offer any statewide computer-delivered testing?

- Yes, we are in our fourth year

What grade and subject-area tests are offered via computer?

- Science grades 4–12 (including physics, chemistry, biology, and earth systems)
- Language Arts grades 3–11
- Math grades 3–12 (including pre-algebra, Elementary Algebra, Intermediate Algebra, Geometry, Applied Math 1 & 2)

Is testing via computer voluntary or mandatory?

- Voluntary

If voluntary, about what percent of testing takes place via computer?

- Last year, 7.3%, this year up to 10%

Approximately how many computer-delivered tests were administered in your state last year (i.e., number of tests delivered via computer)?

- 81,000

Are the computer-delivered tests "fixed" or "scrambled" forms, OR computer-adaptive?

- Fixed, to mirror appearance in Paper/Pencil based testing

Do the computer-delivered tests include short-constructed responses?

- No

Do the computer-delivered tests include extended-writing responses?

- No

How quickly are computer-delivered test scores made available to the student, school, and district (i.e., what is the reporting turnaround time for these different groups)?

- ***Student results:*** Class reports are available 48 hours after every student in the class hierarchy has submitted a completed test, reports are per student.
- ***School results:*** [see below]
- ***District results:*** School and District reports wait until both computer and paper are finished and submitted to the state, and then ran through the same scoring process. Error reports are returned to the districts within 5 days, to be corrected by the districts within 5 days. If no errors or no paper based, School and District reports are returned in 5 days. Otherwise, the timeline varies dependant on error correction.
- ***Statewide results:*** As stated above, timelines vary, but usually 2.5 months after testing has ended, statewide reports could be available.

Which vendor(s) are providing your test engine(s) for your computer-delivered testing?

- PEM [Pearson Education Measurement] providing for 06–07 school year

VIRGINIA

Does your state offer any statewide computer-delivered testing?

- Yes

What grade and subject-area tests are offered via computer?

- We offer all of the Virginia Standards of Learning (SOL) tests in the online environment EXCEPT the direct writing assessment. This includes End-of-Course tests (Algebra I, Plain English Algebra I, Geometry, Algebra II, Earth Science, Biology, Chemistry, English: Reading, VA & US History, World History I, World History II, World Geography), the Grades 3 through 8 Reading & Mathematics tests (including Plain English Mathematics), Grades 3, 5, & 8 Science, and 4 History tests for elementary/middle school students (Virginia Studies, US I, US II, Civics & Economics)

Is testing via computer voluntary or mandatory?

- Voluntary

If voluntary, about what percent of testing takes place via computer?

- We offer all of the Virginia Standards of Learning (SOL) tests in the online environment EXCEPT the direct writing assessment. This includes End-of-Course tests (Algebra I, Plain English Algebra I, Geometry, Algebra II, Earth Science, Biology, Chemistry, English: Reading, VA & US History, World History I, World History II, World Geography), the Grades 3 through 8 Reading & Mathematics tests (including Plain English Mathematics), Grades 3, 5, & 8 Science, and 4 History tests for elementary/middle school students (Virginia Studies, US I, US II, Civics & Economics)

Approximately how many computer-delivered tests were administered in your state last year (i.e., number of tests delivered via computer)?

- Approximately 1.1 million tests were administered online last year.

Are the computer-delivered tests "fixed" or "scrambled" forms, OR computer-adaptive?

- Fixed forms with the exception of embedded field test items

Do the computer-delivered tests include short-constructed responses?

- Not at this time.

Do the computer-delivered tests include extended-writing responses?

- Not at this time, although VA conducted a Gr. 8 English direct writing field test (extended writing response) online for the first time in Spring 2007. This is currently being hand-scored.

How are these responses scored?

- This is currently being hand-scored.

How quickly are computer-delivered test scores made available to the student, school, and district (i.e., what is the reporting turnaround time for these different groups)?

- ***Student results:*** Once post-equating is completed, student-level scores are available to authorized personnel immediately upon the submission of the online test.
- ***School results:*** Once post-equating is completed, school level data extract files are updated nightly during testing so updated student data and score information is available each morning to authorized school and division personnel.

- ***District results:*** Once post-equating is completed, district level data extract files are updated nightly during testing so updated student data and score information is available each morning to authorized district level personnel.
- ***Statewide results:*** once post-equating is completed, state level data extract files are updated nightly during testing so updated student data and score information is available each morning to authorized state level personnel.
- Printed, formatted reports, summary-level data and reports, and finalized student level data files are available to authorized personnel at the respective district and state levels once testing is completed and student demographic data has been reviewed and updated for accuracy.
- For tests that are graduation requirements, (EOC tests and Gr. 8 Mathematics & Reading), VA maintains a supply of test forms each administration that are "previously administered" and therefore are considered "previously post-equated". These forms are administered to those students who are attempting to earn "verified credits" to graduate at the end of the spring semester. Results to these tests are available immediately to students who complete them online as opposed to having to wait for a new form to be post-equated prior to knowing test results.

Which vendor(s) are providing your test engine(s) for your computer-delivered testing?

- Pearson Educational Measurement

Is there anything else you would like to share from your state's experience in implementing computer-delivered testing that might inform the State of South Carolina's Computer-Based/Computer-Adaptive Feasibility Study?

- Virginia implemented online testing using a phased approach (beginning in 2000). (The first operational online administration in 2001 included 3 EOC tests administered in 15 local divisions). VA currently administers online testing in all 132 local divisions and all assessments are offered online currently except the English: Writing (direct writing) assessment. VA purposely partnered between the state divisions of assessment and technology and encouraged/required a similar partnership at the local levels. These partnerships have been one of the critical factors in the successful implementation of Virginia's program both from a financial and an implementation perspective. Virginia has learned many lessons throughout its experience in the areas of implementation, technology, comparability, testing environments, accommodations, security, change management, funding, procurement, etc. I have had the opportunity to manage Virginia's online testing efforts since the inception of the initiative and have the support of the VA Department of Education to share our experiences and information. Virginia is also interested in learning of other states' approaches to the same issues as well as to short and extended constructed response assessments and computer-adaptive assessments. South Carolina and/or DRC may feel free to contact me for additional information if desired.

West Virginia

Does your state offer any statewide computer-delivered testing?

- West Virginia has conducted Online Writing Assessment—Grades 7 and 10 for three years.

What grade and subject-area tests are offered via computer?

- Writing – Grades 7 and 10

Is testing via computer voluntary or mandatory?

- Mandatory

Approximately how many computer-delivered tests were administered in your state last year (i.e., number of tests delivered via computer)?

- Approximately 44,000 students in grades 7 and 10 tested via computer each year in a 4-week window.

Are the computer-delivered tests "fixed" or "scrambled" forms, OR computer-adaptive?

- Descriptive, Narrative, Expository, Persuasive prompts are randomly assigned to students

Do the computer-delivered tests include short-constructed responses?

- Essay response

How are these responses scored?

- Essays scored by artificial intelligence — Vantage Learning

Do the computer-delivered tests include extended-writing responses?

- See above — essay writing

How are these responses scored?

- "Artificial intelligence" (i.e., a computer program scores the responses), see above

How quickly are computer-delivered test scores made available to the student, school, and district (i.e., what is the reporting turnaround time for these different groups)?

- All reports (student, school, district, state) are delivered within 60 days. Students/schools receive reports before the end of the school year.

Which vendor(s) are providing your test engine(s) for your computer-delivered testing?

- CTB/McGraw-Hill is our testing vendor; Vantage Learning is the subvendor who scores the essays.

Is there anything else you would like to share from your state's experience in implementing computer-delivered testing that might inform the State of South Carolina's Computer-Based/Computer-Adaptive Feasibility Study?

- Importance of reliable/valid prompts (field testing, etc.); reliability/dependability of scoring engine.

If your state does not currently offer any statewide computer-delivered testing, is your state planning (or has conducted) any pilots or feasibility studies of computer-delivered testing?

- West Virginia is exploring the idea of Online Writing Assessment for all grades 3–10. Also, WV is planning to combine the WA score with the Reading/Language Arts test and thus WA would be included in the school's AYP calculation.

WYOMING

Does your state offer any statewide computer-delivered testing?

- Yes

What grade and subject-area tests are offered via computer?

- Grades 3–11, math, reading and science

Is testing via computer voluntary or mandatory?

- Required. Only method of delivery available except for Print Accommodation.

Approximately how many computer-delivered tests were administered in your state last year (i.e., number of tests delivered via computer)?

- 61,200

Are the computer-delivered tests "fixed" or "scrambled" forms, OR computer-adaptive?

- Fixed, with multiple forms

Do the computer-delivered tests include short-constructed responses?

- Not this year. They will starting 2008

If yes, in what grades and subjects are these offered?

- Grades 3–11

How are these responses scored?

- "By hand"

Do the computer-delivered tests include extended-writing responses?

- Not yet. They will beginning 2008

If yes, in what grades and subjects are these offered?

- Grades 3–11

How are these responses scored?

- By hand

How quickly are computer-delivered test scores made available to the student, school, and district (i.e., what is the reporting turnaround time for these different groups)?

- ***Student results:*** Multiple choice, beginning second week of testing window. All results four weeks after close of window.
- ***School results:*** As above, with school reports four weeks after close of window
- ***District results:*** District reports four weeks after close of window
- ***Statewide results:*** State reports four weeks after close of the window

Which vendor(s) are providing your test engine(s) for your computer-delivered testing?

- Harcourt Assessment, Inc.

Is there anything else you would like to share from your state's experience in implementing computer-delivered testing that might inform the State of South Carolina's Computer-Based/Computer-Adaptive Feasibility Study?

- Wyoming spent one year ensuring the infrastructure of all schools met bandwidth and computer speed, as well as providing statewide bandwidth. Test tryout of infrastructure with pilot test identified many specific additional issues. Private schools had difficultly meeting testing requirements. Students and schools really like the online testing, but it does impact use of computers and tech resources during the testing window. Wyoming districts are tech ready, and early adopters. They were willing to experience the glitches in order to have more rapid reporting and better control of test security.

COMPONENT 1 – DESCRIPTION OF STATE ASSESSMENT PROGRAMS WHICH ARE *NOT* COMPUTER-BASED OR COMPUTER-ADAPTIVE

ALABAMA

Does your state offer any statewide computer-delivered testing?

- No

If your state does not currently offer any statewide computer-delivered testing, is your state planning (or has conducted) any pilots or feasibility studies of computer-delivered testing?

- No – however, Alabama uses DIBELS for grades K–2 which allows student scores to be entered electronically and reports to be generated for immediate results.

ARIZONA

Does your state offer any state-wide computer-delivered testing?

- Currently we do not offer any assessments online. However, we will be conducting a pilot of 8th grade science this Spring with Pearson Education Measurement.

What grade and subject-area tests are offered via computer?

- Science – Pilot at 8th grade needed to check feasibility of offering online alternative.

Is testing via computer voluntary or mandatory?

- Mandatory field test—A survey of representatively selected schools was given to determine if they had compatible hardware and access.

If voluntary, about what percent of testing takes place via computer?

- N/A. However if successful, we may be adding a voluntary component to assessment.

Approximately how many computer-delivered tests were administered in your state last year (i.e., number of tests delivered via computer)?

- 0

Are the computer-delivered tests "fixed" or "scrambled" forms, OR computer-adaptive?

- The pilot will have fixed multiple forms; future assessments have not been established or planned until success of pilot.

Do the computer-delivered tests include short-constructed responses?

- No

Do the computer-delivered tests include extended-writing responses?

- No

How quickly are computer-delivered test scores made available to the student, school, and district (i.e., what is the reporting turnaround time for these different groups)?

- Since the purpose of the pilot is to field test items, no reporting will be available.

Which vendor(s) are providing your test engine(s) for your computer-delivered testing?

- Pearson [Educational Measurement] will be conducting the pilot.

Is there anything else you would like to share from your state's experience in implementing computer-delivered testing that might inform the State of South Carolina's Computer-Based/Computer-Adaptive Feasibility Study?

- It would be helpful if South Carolina and DRC would share the results of this survey with the states responding to it. For our part, it would be helpful and necessary in our requests to the legislature for funding purposes.

If your state does not currently offer any statewide computer-delivered testing, is your state planning (or has conducted) any pilots or feasibility studies of computer-delivered testing?

- Yes, all answers above pertain to pilot to be conducted this Spring. The expansion to online needs to be implemented with care in order to capitalize on successful ventures and prevent disaster.

ARKANSAS

Does your state offer any statewide computer-delivered testing?

- No

If your state does not currently offer any statewide computer-delivered testing, is your state planning (or has conducted) any pilots or feasibility studies of computer-delivered testing?

- No

California

Does your state offer any statewide computer-delivered testing?

- No

If your state does not currently offer any statewide computer-delivered testing, is your state planning (or has conducted) any pilots or feasibility studies of computer-delivered testing?

- The contractor for the California Standards Tests is conducting surveys to determine the technical capabilities of school districts to offer computer-based testing. The contractor will report to the California Department of Education (CDE) and the State Board of Education (SBE). Implementation of pilot studies or further development of computer-based testing will only proceed if approved by the CDE and SBE.

Colorado

Does your state offer any statewide computer-delivered testing?

- No

If your state does not currently offer any statewide computer-delivered testing, is your state planning (or has conducted) any pilots or feasibility studies of computer-delivered testing?

- There were a few re: writing, but this is a long-term capacity process.

DELAWARE

Does your state offer any statewide computer-delivered testing?

- No, we do not current offer any CBT. We did a small-scale pilot a few years ago but are not actively doing any statewide CBT.

If your state does not currently offer any statewide computer-delivered testing, is your state planning (or has conducted) any pilots or feasibility studies of computer-delivered testing?

- Yes, we will be planning pilots and feasibility studies under a new contract. The Request for Proposals is currently posted on our web site (www.doe.k12.de.us/rfplisting). The new contract for the annual summative assessment will start on or about July 1, 2007 and includes a section on transitioning to online assessments. The new contract will have the first operational assessment in spring 2009 although the transition to online assessment will follow a different timeline.

HAWAII

Does your state offer any statewide computer-delivered testing?

- Hawaii does not do any computer-based testing.

If your state does not currently offer any statewide computer-delivered testing, is your state planning (or has conducted) any pilots or feasibility studies of computer-delivered testing?

- [No response]

ILLINOIS

Does your state offer any statewide computer-delivered testing?

- No

If your state does not currently offer any statewide computer-delivered testing, is your state planning (or has conducted) any pilots or feasibility studies of computer-delivered testing?

- Illinois will conduct a comparability of computer to paper study in 2007 and 2008.

IOWA

Does your state offer any statewide computer-delivered testing?

- No

If your state does not currently offer any statewide computer-delivered testing, is your state planning (or has conducted) any pilots or feasibility studies of computer-delivered testing?

- The Iowa Testing Program at the University of Iowa has responsibility for the statewide testing. They have discussed moving to a computer-delivered testing program, although no specific timeline or plans have been proposed to the Iowa Department of Education.

LOUISIANA

Does your state offer any statewide computer-delivered testing?

- No

If your state does not currently offer any statewide computer-delivered testing, is your state planning (or has conducted) any pilots or feasibility studies of computer-delivered testing?

- The Louisiana Department of Education is conducting an online field test of an Algebra I end-of-course test in May 2007, as recommended by Governor Blanco's High School Redesign Commission. This field test will only be available online. Information on vendors is public record; the vendor for this project is Pacific Metrics.

MAINE

Does your state offer any statewide computer-delivered testing?

- No. [But] We are also interested in doing this…A majority of our schools do use NWEA and love it.

If your state does not currently offer any statewide computer-delivered testing, is your state planning (or has conducted) any pilots or feasibility studies of computer-delivered testing?

- We are planning to include it in our RFP for the state assessment. We use the SAT at the 11th grade so we won't have it there, but want to have it elsewhere.

MASSACHUSETTS

Does your state offer any statewide computer-delivered testing?

- No

If your state does not currently offer any statewide computer-delivered testing, is your state planning (or has conducted) any pilots or feasibility studies of computer-delivered testing?

- The Massachusetts Department of Education commissioned a small study in 2005–06 to explore the feasibility of moving the MCAS testing program online. Since then, the Department has met with an external advisory group to consider options for beginning operational testing online in one or more areas beginning in 2008. However, state funding for FY '08 is not expected to cover requisite costs for this initiative; online testing plans are now on hold.

MICHIGAN

Does your state offer any statewide computer-delivered testing?

- We have piloted computer-delivered assessment and are planning on expanding but at the current time, no administration is planned. [see below]

If your state does not currently offer any statewide computer-delivered testing, is your state planning (or has conducted) any pilots or feasibility studies of computer-delivered testing?

- Yes, pilot study from 2005. We tested grade 6 English Language Arts and Social Studies, approx. 2500 students from across the state. All students were part of a wireless grant project using one-on-one wireless laptops. We did score constructed responses by computer as well as by hand.

We are revising this plan to include these possible assessment areas:

- MEAP subject matter tests: current feasibility plan for grade 9 Social Studies, then add grade 6 Social Studies, followed by Grade 5 and 8 Science over the next 6 years.
- Formative assessment online being considered, at this point similar to TEA, in the planning stage only.
- End-of-course assessment - in this state also known as Secondary Credit Assessment. Modularized HS assessment for 17 courses, in the planning stage only.
- Exploring an Online ELPA screening test with Universities as partners.
- Tentative plan for ELPA screening to take place this fall.

MISSOURI

Does your state offer any statewide computer-delivered testing?

- No

If your state does not currently offer any statewide computer-delivered testing, is your state planning (or has conducted) any pilots or feasibility studies of computer-delivered testing?

- We are piloting an online Personal Finance end-of-course exam in the spring of 2007. MCCE [our contractor], is a part of the University of Central Missouri and is hosting and building the software.

MONTANA

Does your state offer any statewide computer-delivered testing?

- No

If your state does not currently offer any statewide computer-delivered testing, is your state planning (or has conducted) any pilots or feasibility studies of computer-delivered testing?

- No

NEBRASKA

Does your state offer any statewide computer-delivered testing?

- No

If your state does not currently offer any statewide computer-delivered testing, is your state planning (or has conducted) any pilots or feasibility studies of computer-delivered testing?

- No

NEVADA

Does your state offer any statewide computer-delivered testing?

- No. Nevada does not have statewide computer-delivered testing at this point. State mandated tests are 'paper and pencil'.

If your state does not currently offer any statewide computer-delivered testing, is your state planning (or has conducted) any pilots or feasibility studies of computer-delivered testing?

- We're in the preliminary planning stages to pilot our High School Proficiency Examination online. We'll probably try it at one or two small districts/schools to start, most likely next fall but perhaps this summer.
- "Measured Progress" (measuredprogress.org) is our main testing contractor for our HSPE. We would probably be piloting their "iTest" software. They have a subcontract with WestEd (wested.org - we asked for this in our original RFP) to do test item and form development.

NEW HAMPSHIRE

Does your state offer any statewide computer-delivered testing?

- No — not at this time.

If your state does not currently offer any statewide computer-delivered testing, is your state planning (or has conducted) any pilots or feasibility studies of computer-delivered testing?

- No. We have piloted a computer delivered read aloud accommodation for the mathematics test at grade 10 in May 2006, and we are looking at adopting such an accommodation for the full assessment system in the near future. But, we are not looking at any computer platform for the delivery of state assessments now or in the near future. (Too bad, at least for writing...)

- We have not liked any of the commercial versions [of read-aloud software] - the computer-generated voice is not good enough! We are working with two people from Boston College on a platform designed for state assessment with a human voice. It worked VERY WELL with grade 10 students when we tried it.
- Last May, when we piloted the product, we increased to 500 the number of students making use of the read aloud accommodation in mathematics (up from 50 or so the year before). Much of the increase was with ELLs.

NEW JERSEY

Does your state offer any statewide computer-delivered testing?

- No

If your state does not currently offer any statewide computer-delivered testing, is your state planning (or has conducted) any pilots or feasibility studies of computer-delivered testing? (If yes, please explain.)

- Yes, we expecting to transition to computer-based testing in the coming years, and our recent RFP for grades 3–8 addresses that goal, but we have no specific timetable as yet for achieving this goal.

NEW MEXICO

Does your state offer any statewide computer-delivered testing?

- No

If your state does not currently offer any statewide computer-delivered testing, is your state planning (or has conducted) any pilots or feasibility studies of computer-delivered testing? (If yes, please explain.)

- No

New York

Does your state offer any state-wide computer-delivered testing?

- No

If your state does not currently offer any statewide computer-delivered testing, is your state planning (or has conducted) any pilots or feasibility studies of computer-delivered testing?

- We are researching but have not put specific plans or pilots into effect.

North Dakota

Does your state offer any statewide computer-delivered testing?

- No, North Dakota does not have statewide computer-delivered testing. A large number of our school districts independently contract with NWEA for MAP testing, however.

If your state does not currently offer any statewide computer-delivered testing, is your state planning (or has conducted) any pilots or feasibility studies of computer-delivered testing?

- There is nothing definite at this time.

OHIO

Does your state offer any statewide computer-delivered testing?

- No

If your state does not currently offer any statewide computer-delivered testing, is your state planning (or has conducted) any pilots or feasibility studies of computer-delivered testing?

- No. The only thing we currently plan to do is to put the practice tests online.

TENNESSEE

Does your state offer any statewide computer-delivered testing?

- No

If your state does not currently offer any statewide computer-delivered testing, is your state planning (or has conducted) any pilots or feasibility studies of computer-delivered testing?

- Tennessee has piloted online assessments for our Math Foundations II and Algebra I courses.
- The vendor was Pearson Educational Measurement/ ETS. We are still looking at the comparability study and have not made any decisions at this time. These are end of course assessments, typically 9th and 10th grade courses. Again, no decisions have been made as far as replacing the paper-pencil test.

PENNSYLVANIA

Does your state offer any statewide computer-delivered testing?

- No

If your state does not currently offer any statewide computer-delivered testing, is your state planning (or has conducted) any pilots or feasibility studies of computer-delivered testing?

- No, we have not conducted any pilots. We have talked about computer delivered tests but only in terms that someday we feel sure we will need to do it.

RHODE ISLAND

Does your state offer any statewide computer-delivered testing?

- No

If your state does not currently offer any statewide computer-delivered testing, is your state planning (or has conducted) any pilots or feasibility studies of computer-delivered testing?

- Possibly through our partnership with Achieve. The multi-state end-of-course Algebra II Exam may have a computer based option.

VERMONT

Does your state offer any statewide computer-delivered testing?

- No

If your state does not currently offer any statewide computer-delivered testing, is your state planning (or has conducted) any pilots or feasibility studies of computer-delivered testing?

- We are very interested in computer-delivered testing, but our issue is capacity. We still have many schools with dial-up connections, and insufficient terminals to make web-based testing feasible.

WASHINGTON

Does your state offer any state-wide computer-delivered testing?

- No

If your state does not currently offer any statewide computer-delivered testing, is your state planning (or has conducted) any pilots or feasibility studies of computer-delivered testing?

- We are exploring options for pilots and comparability studies, however, we do not have specific dates to implement computer delivered testing at this point.

WISCONSIN

Does your state offer any statewide computer-delivered testing?

- No, not at this time.

If your state does not currently offer any statewide computer-delivered testing, is your state planning (or has conducted) any pilots or feasibility studies of computer-delivered testing

- No, none at this time.

COMPONENT 1 – REFERENCES AND LINKS

Abell, M., and Lewis, P., "Universal Design for Learning: A Statewide Improvement Model for Academic Success" *Information Technology and Disabilities E-Journal*, XI, 1 (August 2005). Available at http://www.rit.edu/~easi/itd/itdv11n1/abell.htm

Associated Press, "Online math test stymies ESL students," *Post-Bulletin* (Rochester, MN), April 27, 2007. Available at http://www.postbulletin.com/newsmanager/templates/localnews_story.asp?a=292097&z=2)

Bakia, M., Mitchell, K., and Yang, E. (2007). *State Strategies and Practices for Educational Technology: Volume I—Examining the Enhancing Education through Technology Program*, U.S. Department of Education. Available at http://www.ed.gov/rschstat/eval/tech/netts/netts-vol1.pdf

Core Curriculum Content Standards, Technological Literacy (2004), New Jersey Department of Education. Available at http://education.state.nj.us/cccs/?_list_cpi;c=8;s=1;g=4

Gold, M. (2007) "Firm Pledges to Fix Online Exam Glitches," *Washington Post* (Washington, D.C.) June 5, 2007.) Available at: http://www.washingtonpost.com/wp-dyn/content/article/2007/06/04/AR2007060401643.html

New England Compact Enhanced Assessment Project, "Computer-Based Read-Aloud Tools: An Effective Way to Accommodate Students with Disabilities," Education Development Center, Inc. (February 2005). Available at http://www.necompact.org/NECompact%20Summaries_0205.pdf

Standard Course of Study, Computer/Technology Skills (n.d.), North Carolina Department of Public Instruction. Available at http://www.dpi.state.nc.us/curriculum/computerskills/scos/10grade3

Standards of Learning for Computer Technology (2005), Virginia Department of Education. Available at http://www.pen.k12.va.us/VDOE/Superintendent/Sols/compteck12.doc

"Technology Counts 2007: A Digital Decade," *Education Week* (Vol. 26, Issue 30) March 29, 2007. Available at http://www.edweek.org/ew/toc/2007/03/29/index.html

Texas Essential Knowledge and Skills, Technology Applications (1998), Texas Education Agency. Available at http://www.tea.state.tx.us/rules/tac/chapter126/ch126a.html#126.2

Thompson, S., Thurlow, M., and Moore, M. "Using Computer-based Tests with Students with Disabilities" (Policy Directions, Number 15, January 2003, National Center on Educational Outcomes). Available at http://www.education.umn.edu/nceo/OnlinePubs/Policy15.htm

Williams, A. "It's back to pencil, paper for state tests," *The Register-Guard* (Eugene, OR, May 7, 2007. Available at http://www.registerguard.com/news/2007/05/07/c1.cr.testing.0507.p1.php?section=cityregion

COMPONENT 2 – A REVIEW OF THE LITERATURE ON THE COMPARABILITY OF SCORES OBTAINED BY EXAMINEES WHEN ASSESSMENTS ARE ADMINISTERED BY COMPUTER RATHER THAN PAPER AND PENCIL

The Role and Importance of Comparability Studies

The current computer technologies with high speeds, large storage capacities, general availability, and rapidly shrinking costs make electronic testing (eTesting) an appealing alternative to the traditional paper-and-pencil (P&P) measures. A survey of state departments of education found that twelve states were using some form of online testing in 2003 with half of these states in the piloting phase (EdWeek, 2003b). By the end of 2005, the number of states piloting or offering online tests had grown to 25 (EdWeek, 2005). Based on the report of the U.S. Department of Education, twenty-six states reported either offering technology-based academic assessment or funding research and development activities that supported student assessment in FY 2003 (U.S. Department of Education, 2007, p.19). By 2008, almost all states are expected to be using some form of online testing (Williams, n.d.).

The central role of large-scale, state-mandated tests in school accountability makes it essential that any transition to eTesting be seamless. Continuity of AYP determinations cannot be ensured if the measurement changes significantly. Although this fact of life means any discontinuity in the measurement will present problems for state departments of education (SDE) and the districts, eTesting is probably inevitable. Thus, the inevitable disruptions that eTesting may cause must be anticipated and mitigated.

For those K–12 agencies that plan to offer the same tests on both paper and computer, comparability will be a concern because of the familiarity and pervasiveness of P&P testing and because of the requirements of NCLB for continuity and fairness. Without those influences, “comparability is a short lived phenomenon. Five years from now, establishing the comparability of computer-based assessment to the *outmoded* paper-based will be a non-issue. No one starting from scratch today would even consider building a paper-based assessment.”[1]

One important consideration in determining the feasibility of electronic testing is the *interchangeability* of scores between eTesting and P&P. *Interchangeable* means that one does not need to know which testing mode was used in order to interpret the score. Researchers have been attempting to answer this question as long as there have been computer-administered tests.

Score Comparability Studies fundamentally are *validity* studies. Because of the necessity of continuity with previous assessments and because the transition to eTesting from paper-and-pencil based tests (P&P) rarely happens in one fell swoop, the P&P version is taken to be the gold standard for validity. Under these rules, eTesting is appropriate if it produces scores that are interchangeable with scores from P&P.

“When interpreting scores from the computerized versions of conventional tests, the equivalence of scores from computerized versions should be established and documented before using norms or cut scores obtained from conventional tests.” (American Psychological Association, 1986, p. 18). The *Standards for Educational and Psychological Testing* (AERA, 1999) recommends empirical validation of the computerized versions of tests: “A clear rationale and supporting evidence should be provided for any claim that scores earned on different forms of a test may be used interchangeably.” (AERA, 1999, p. 57). *The Association of Test Publishers* provide more specific guidance for eTesting. Standard 2.10 states, “Developers of computer-based tests should consider how aspects of computer delivery might impact fairness and equity and take appropriate action to minimize their effect. These factors may include aspects of test design, content, specific items, or format elements.” (ATP, 2002, p. 18).

[1] Poggio, J. (2007). South Carolina expert panel on computer-based testing discussion. March 28, 2007. Appendix B, p B-5.

Kolen (1999-2000) summarizes four possible sources of differences in scores for eTesting versus P&P:

1. test questions,
2. test scoring,
3. testing conditions, and
4. examinee groups.

Comparability studies also explore other possible effects, such as test-taking strategies and interaction between the testing mode and content area.

Even within eTesting, there are comparability issues (Bennett, 2002a). Among schools (and even within the same school), monitor size, screen resolution, keyboard layout, connection speed, and other technical characteristics may vary, affecting the manner in which the student experiences the item and the test. Any of these variations may affect scores unfairly. Brideman, Lennon, and Jackenthal (2001) found that using smaller screens, and thus increasing the need for scrolling, diminished test performance on reading comprehension items by a small but nontrivial amount. Similarly, Powers and Potenza (1996) presented evidence to suggest that essays written on laptop might not be comparable to those written on desktops with more convenient keyboards and screen displays.

Differences in computer familiarity and anxiety may lead to differences in performance (Bennett, 2002a). Although physical access to school computers differs little by income and racial group, home-access disparities are still substantial (U.S. Department of Commerce, 2002).

Purposes and Scope of the Literature Review and Evaluation

This review will focus on studies of educational achievement tests conducted on students in grades K-12 and that were published or presented in 1997 or later. Overall comparability results will be presented two ways: based on the judgment of the original authors and by a distribution of effect sizes.

This review and evaluation begins with a discussion of the challenges of the literature review because of inconsistent findings and varied methodologies applied in the studies. The meta-analysis method with evaluation of effect size to evaluate literature is described. This approach divides comparability studies into two categories: multiple-choice-only tests and constructed response or writing assessments.

CHALLENGES AND INCONSISTENCIES OF COMPARABILITY STUDIES

While the preponderance of the evidence suggests that, for multiple-choice-only tests, student performance is not significantly different for different modes of administration, some studies suggest students might do better on computer and others suggest they might do better on paper. One must consider the stakes associated with the test when determining practical significance of the difference due to the mode of administration and in reaching consensus about how much is too much.

Many comparability studies found computer tests to be equivalent in difficulty or slightly easier than paper tests (Bridgeman, Bejar, & Friedman, 1999; Choi & Tinkler, 2002; Mead & Drasgow, 1993; Pearson Educational Measurement, 2002, 2003; Poggio, Glasnapp, Yang, and Poggio, 2005; Pommerich, 2004; Pomplun, Frey & Becker, 2002; Russell, 1999; Russell & Haney, 1997; 2000; Russell & Plati, 2001; Schaeffer, Bridgeman, Golub-Smith, Lewis, Potenza, & Steffen, 1998; Nichols & Kirkpatrick, 2005; Taylor, Jamieson, Eignor, & Kirsch, 1998;

Zandvliet & Farragher, 1997; Wang, 2004). Student performance was similar across demographics (gender, academic placement, and SES).

Pommerich (2004) analyzed Grade 11 and 12 students on a computerized fixed-form test in English, reading, and science reasoning in 1998 and 2000. She found student scores were about one raw score higher or lower between computer versions and paper version across subjects. Using the National Assessment of Educational Progress (NAEP) writing tests for Grade 8 students, Bennett (2003) found scores were not significantly different. Studying performance on the Stanford Diagnostic Reading and Mathematics Tests, 4th edition (SDRT 4 and SDMT 4) for students in grades 2–12, Wang found that, overall, there were no significant differences in total test score means based on administration mode, mode order, or mode-by-mode order interactions. Some studies found that students tend to write more on computer tests than on paper tests, but the writing is not necessarily better (Nichols, 1996; Russell, 1999; Russell & Haney, 1997).

On the other hand, some K-12 studies found students performed poorer on computer tests than paper tests (Cerrillo & Davis, 2004; O'Malley, Kirkpatrick, Sherwood, Burdick, Hsieh, & Sanford, 2005; Russell & Plati, 2001). Choi and Tinkler (2002) found student scores from the computer tests in reading for grades 3 through 10 math and reading of the Oregon state assessment were lower than the paper tests, especially for third-graders. Also, they found that computer familiarity was related to computer test performance; students who rarely used a computer tended to perform poorer in both math and reading than those students who had more computer experience. In the Virginia state assessment, students scored one raw score point higher on the paper test in both English and earth science (Pearson Educational Measurement, 2002; 2003).

Finally, inconsistent findings in the mode effects have been found on constructed response questions as well as for multiple-choice only tests. Russell (1999) compared Grade 8 student scores on constructed-response questions across subjects. He found that there were no significant differences of scores in language arts and math, but students scored significantly higher in science for electronic tests. Also, Russell (2002) and Russell and Haney (1997, 2000) found that middle-school students performed similarly on paper tests and computer tests for multiple-choice questions but significantly better on computer tests for science and language arts with constructed-response questions.

There are a number of possible factors that might help explain some of this inconsistency. First, not all computerized test administration systems are the same. Software developers have learned a great deal over the years about how to make test administration software more user-friendly and less obstructive. For instance, some early systems required students to scroll both up and down *and* side to side to view an entire passage or large item. Newer systems recognize that this makes an item harder and minimize scrolling to a single dimension at most and try to avoid scrolling entirely.

Also, students expect to have the same aids and flexibility that they were accustomed to on P&P when taking an eTest. For example, the inability to highlight text or to cross off incorrect responses led to high levels of student frustration in the first implementations, which has been changed in later systems.

The flexibility to review and edit responses to earlier items has also been found to be important to students (Stocking, 1997). Students report lower levels of satisfaction with systems that lack this capability. Studies have found slightly higher test scores when students are permitted to return to previous items, although the studies have not addressed the question of whether the improved performance was due to elimination of careless errors or was based on information gleaned from intermediate items (Vispoel, Hendickson, & Bleiler, 2000).

Second, the strongest, and most informative, type of research study is one in which students are randomly assigned to testing conditions with similar incentives to do well regardless of which test mode they take. But it is often impractical to have half the students in a classroom taking a test on computer while the other half are taking the test on paper. Sometimes it is easier to let students (or teachers) decide which administration mode they will use. Results from this latter kind of study might be affected by the sampling effects of who has chosen to participate in each group. For example, did better students have more familiarity with computers and thus choose to participate in the online group?

Methodologies of Comparability Studies

A majority of the comparability research has focused on the differences in means and standard deviations of test scores (Makiney, Rosen, Davis, Tinios & Young 2003; Mead & Drasgow, 1993; Merten, 1996; Pinsoneault, 1996). Also, classical item and form analyses such as p-values, response-option (distracter) analyses, item-to-total-score correlations, item difficulty values, item-model fit measures, standard errors of measurements, test form reliability, summary descriptive statistics, frequency distributions and percentile rank information, have been used for comparability studies. The comparison of person-fit statistics can also be made across modes and can be disaggregated by various breakout groups including ethnicity and gender. In some, the root mean-squared difference (RMSD) statistic was used to compare performance across testing modes (Paek, 2005; Raju, Laffitte, & Byrne, 2002; Donovan, Drasgow, & Probst, 2000; Pinsoneault, 1996).

In 2003, Makiney, et al. proposed two common techniques for comparability studies. The first one was the structural equation modeling method involving confirmatory factor analysis. It can be used for tests of invariances and analyzing multiple latent constructs and multiple populations simultaneously. The second one was based on item response theory (Raju, Laffitte, & Byrne, 2002). Two IRT analyses can be used to test for the equivalency of item parameters from two modes, the differential item functioning analysis (DIF) and the differential test functioning (DTF).

Different comparability studies employed different methodologies. Careful review of the literature shows that author judgment of comparability was subjective and inconsistent across studies. Other studies (e.g., Poggio, et al. 2005b; Wang, Young, & Brooks, 2004; Yang, Poggio, Glasnapp & Poggio, 2007) made no evaluative comments. To achieve consistency for this review, effect sizes were calculated as:

1. $$ES = \frac{\bar{x}_{computer} - \bar{x}_{paper}}{\sqrt{\dfrac{(n_{computer} - 1)s^2_{computer} + (n_{paper} - 1)s^2_{paper}}{n_{computer} + n_{paper} - 2}}}.$$

Across all the studies, effect sizes (or the information necessary to calculate the effect sizes) were presented for 85 content area grade combinations. Contact with the authors of other papers allowed the calculation of four other effect sizes for a total of 89. Sample sizes for these studies ranged between 42 and 4,333 students.

For all analyses, a positive difference indicates students performed better when tested on a computer and a negative difference indicates students performed better on paper.

Comparability for K–12 Students for Multiple-Choice Only Tests

Table 2.1
Recent K–12 Comparability Studies of Multiple-Choice Tests
(**Boldface** indicates the entry was included in the effect size analysis)

	More Difficult Administration Mode		
	Computer	**Comparable**	**Paper**
Math	**Choi & Tinkler (2002), G3** Cerillo & Davis (2004), Algebra Sandene, Bennett, Braswell, & Oranje (2005), Way, Davis, & Fitzpatrick (2006), G11	**Kim & Hunyh (2006), Algebra** **Kingston (2002), G1,4,6,8** PEM (2002), Algebra PEM (2003), Algebra II Nichols and Kirkpatrick (2005) **Poggio, Glassnapp, Yang, & Poggio (2005), G7** Russell (1999), G8 Russell (2002), G6,7,8 **Russell & Haney (1997**, 2002), G6,7,8 Wang (2004), G2–5,7–12 Way, Davis, & Fitzpatrick (2006), G8	**Choi & Tinkler (2002), G10**
Language Arts		**Kim & Huynh (2006), HS** **Kingston (2002), G1,4,6,8** **Pommerich (2004), G11–12** Russell (1998), G8 Russell (2002), G6,7,8 **Russell & Haney (1997,** 2002), G6,7,8 Way, Davis, & Fitzpatrick (2006), G11	Russell (2002), G6,7,8 Russell & Haney (1997, 2002), G6,7,8

Table 2.1 (continued)
Recent K–12 Comparability Studies of Multiple-Choice Tests
(**Boldface** indicates the entry was included in the effect size analysis)

More Difficult Administration Mode			
	Computer	**Comparable**	**Paper**
Reading	**Choi & Tinkler (2002), G3** Cerillo & Davis (2004), HS English Way, Davis, & Fitzpatrick (2006), G8	**Kingston (2002), G1,4,6,8** Nichols & Kirkpatrick (2005), PEM (2002), HS English **Pommerich (2004), G11,12** Russell (1999), G8 Russell (2002), G6,7,8 Russell & Haney (1997, 2002), G6,7,8 Wang (2004), G2–5,7–12	**Choi & Tinkler (2002), G10** **Higgins, Russell, & Hoffman, 2005, G4** O'Malley, et al. (2005), G2–5,8
Science	Cerillo & Davis (2004), Biology	**Kim & Huynh (2006), Physical Science, Biology** **Kingston (2002), G4,6,8** PEM (2002), Earth Science PEM (2002), Biology **Pommerich (2004), G11,12** Russell (1998), G8 Russell (2002), G6,7,8 Way, Davis, & Fitzpatrick (2006), G11	Russell (1999), G8 Russell (2002), G6,7,8 Russell & Haney (1997, 2002), G6,7,8
Social Studies		**Kingston (2002), G4,6,8** Way, Davis, & Fitzpatrick (2006), G8,11	

For most studies of the comparability of computer and paper administration, the tests have consisted of multiple-choice items. Paek (2005) presented a table summarizing the results of recent comparability studies. Table 2.1 is adapted from hers and updated to include more recent studies. The judgments of comparability that are first reported here are the judgments made by the original authors themselves. Totaling all the number of grade and content area combinations, the results were deemed comparable in 79 of 108 cases. The students scored higher on the computer-administered test in 21, and higher on the paper test in 8.

The effect size is used to determine if the differences are sufficiently large to reveal real differences across studies. The mean effect size is 0.02, which says, across these studies, students performed 0.02 standard deviations better on a computer-administered test than on a paper-administered test. Table 2.2 presents a stem-and-leaf plot of the effect sizes.

Table 2.2
Stem-and-Leaf Plot of 50 Effect Sizes

-.4	1
-.3	0
-.2	533
-.1	755410
-0	99988766665444433222221
0	0000000001111222333334444456668999
.1	001123344567789
.2	5558
.3	1
.4	1

A stem-and-leaf plot combines the characteristics of a table and a histogram. For a given row, the first column represents the first digit of the effect size and the numbers in the second column represent the second digit of each effect size in that category. Thus, the third row of

Table 2.2 indicates three studies, one with an effect size of -0.25 and two with an effect size of -0.23.

Table 2.3 presents the mean differences broken down by content area and grade.

Table 2.3
Mean Effect Size by Content Area and Grade

Grades	Content Area					
	Math	Reading	ELA	Science	Social Studies	All
Elementary	-0.01	-0.05	-0.03	0.03	0.14	-0.02
Middle	-0.04	0.02	0.06	0.13	0.17	0.02
High	-0.09	0.09	0.04	0.06	--	0.03
All	-0.05	0.05	0.03	0.08	0.16	0.02

Table 2.4 presents the number of studies that form the basis of each mean in Table 2.3.

Table 2.4
Number of Studies by Content Area and Grade

Grades	Content Area					
	Math	Reading	ELA	Science	Social Studies	All
Elementary	9	7	2	1	1	20
Middle	10	9	3	3	1	26
High	12	22	4	5	0	43
All	31	38	9	9	2	89

The mean effect sizes from the 89 studies that went into Table 2.3 were compared using the generalized linear model[2]. The 0.02 overall mean difference between computer and paper was statistically significant at the 0.024 level, indicating a small but likely replicable advantage for students taking tests on computer. The main effect for content area was statistically significant at the 0.008 level. Specifically, the confidence intervals for the estimated mean effect size for mathematics and social studies did not overlap. The main effect for grade was not statistically significant at any commonly accepted level, nor was the interaction between grade and content area.

If the change is real and can be described simply by the effect size, the difference between modes can be equated out. That is, an adjustment to one mode can be made to make the two modes comparable. If the composite trait measured by a particular test changes in ways more complex than a simple mean shift—a condition that can often be detected using any variety of multi-trait, multi-method (MTMM) techniques based on structural equation modeling using conditional differential test functioning (Donovan, Drasgow, and Probst, 2000).

Comparability for K–12 Students on Open-Ended Tests

Far fewer comparability studies have been conducted on open-ended tests and those studies have idiosyncratic designs that make it difficult or inappropriate to look at average effect sizes. Instead, this section presents the highlights of several studies that together suggest that for other open-ended test questions, there is insufficient evidence to determine if students are advantaged or disadvantaged taking tests on computer.

In a 1999 study, Russell found essentially no difference for language arts open-ended items (effect sizes of 0.00 and -0.04 for two sets of three items), a difference favoring paper-and-pencil for mathematics (-0.35 effect size), and a difference favoring computer for science (0.55 effect size). All of these studies were based on small random samples (40-70 per administration condition) with each group administered the test either in paper or computer mode.

[2] The generalized linear model is similar to an analysis of variance but uses maximum likelihood estimation and likelihood ratio chi square tests rather than least squares estimation and F-tests.

A study conducted for Michigan (Pearson Educational Measurement, 2006) compared social studies results of 1,095 grade 6 students who took the state assessment online with a matched group sampled from 112,729 students who took the test on paper. Similarly, they looked at reading and writing results of 1,133 students who took those tests online with a matched group drawn from a total group of 42,872 students who took those tests on paper. The paper does not present effect sizes but does conclude that, for the one reading constructed response item and two writing constructed response items, students performed better on the paper version.

In 2005, Sandene, Horkay, Bennett, Allen, Kaplan & Oranje conducted a comparability study of the National Assessment of Educational Progress grade 8 mathematics and grade 8 writing assessments. For each administration mode, about one thousand students were administered multiple-choice, short constructed response, and extended constructed-response mathematics items[3]. Three of the eight short constructed response items and one of the two extended constructed response items had to be revised significantly to allow them to be administered on the computer. Of the remaining six items, four were statistically significantly more difficult when administered on computer and two were not. The differences in proportion of possible points achieved were -0.17, -0.16, -0.08, -0.08, -0.02, and 0.00.

Comparability for K–12 Students on Direct Writing Assessments

Far fewer comparability studies have been conducted on direct writing tests and those studies have idiosyncratic designs that make it difficult or inappropriate to look at average effect sizes. Instead this section presents the highlights of several studies that together suggest that for direct measures of writing, it appears many students gain a small but significant advantage if

[3] Results of this study were not included in the effect size analysis earlier in this paper because the report did not provide the data necessary to calculate effect sizes. The study did conclude that multiple-choice items were statistically significantly more difficult when administered on computer, though for most items the differences were not large.

they can take the test on computer and if they have received appropriate instruction via computer.

In a study of 30 sixth grade students in one private school (Nichols, 1996) students responded to two writing prompts in counterbalanced order and average scores were 0.19 standard deviations higher for the essay they composed on a word processor. Their responses were also significantly longer (255 words versus 146 words) on the word-processed essay.

In 1997, Russell and Haney found an effect size of 0.9 favoring computers and hypothesized that students accustomed to writing with a word processor were disadvantaged if they had to respond via paper and pencil. Russell and Plati (2001) found eighth and tenth grade students did better on a direct writing measure when administered on computer (effect sizes of 0.71 and 0.51 respectively).

In 2003, Goldberg, Russell and Cook conducted a meta-analysis of studies from 1992 to 2002 regarding the effect of computers on student writing. This study included 26 studies conducted between 1992-2002 focused on the comparison between K–12 students writing with computers versus paper-and pencil. Significant mean effect sizes in favor of computers were found for both the quantity and quality of writing. Studies that focused on revision behaviors revealed mixed results.

Other studies collected for the meta-analysis which did not meet the statistical criteria were also reviewed briefly. These articles indicate that the writing process is more collaborative, iterative, and social in computer classrooms compared to paper-and-pencil environments. The results of meta-analyses suggests that, on average, students who use computers when learning to write are more engaged and motivated in their writing, and they produce written work that is of greater length and higher quality.

In an unpublished study conducted in 2004, Kingston reported on a statewide assessment of direct writing administered to eighth grade students. Schools chose to test using paper or computer administration. Of the 16,596 total students, 14,780, students responded to the test using paper and 1,816 using computer. Both groups took a paper version of a reading test. The group that took the writing test via computer had a mean score 0.32 standard deviations higher than the group that took the writing measure on paper, but their reading scores (which both groups took on paper) were only 0.16 standard deviations higher. The effect size due to computer administration predicted from the reading scores was 0.21, but was greater for the highest scoring examinees (0.35 effect size) and smaller for the lowest scoring examinees (0.07 effect size).

In 2005, Sandene et al. conducted a comparability study of the National Assessment of Educational Progress grade 8 writing assessment. This study found no significant difference in essay score or essay length. The New England Compact (2005) found that students with good computer skills performed better when given direct writing assessments on computer. Students with poor computer skills performed equally in either mode. Its recommendation is that all students be assessed in writing via computer because none are harmed and some are helped.

Test Characteristics

Because of the inconsistency in the findings for all item types, it is appropriate to examine more specific characteristics of the tests in an attempt to isolate the sources of the inconsistencies. Relevant test characteristics include differences in the test content across modes, differences in item and reference material formatting and accessibility, and differences in administration setting and timing.

Test Content. Although the computers are capable of presenting many more items and more flexible formatting than paper and pencil and can test different skills, most test developers have attempted to avoid the content equivalence issue by using identical types of questions in both testing modes.

Long Passages. Long reading passages on a computer tend to be more difficult than on paper (Murphy, Long, Holleran, & Esterly, 2000; O'Malley, et al., 2005). Researchers have speculated that the computer interface interferes with students' comprehension strategies, preventing them from underlining text and from using visual cues to locate information within a passage.

Scrolling seems to negatively affect students' test scores in passage-based assessments (Choi & Tinkler, 2002; Pommerich, 2004).

Although the hypothesis did not reach statistical significance, there is evidence of modal differences in student performance based on their computer skills and knowledge (Russell & Hoffmann, 2005, p. 31). Russell and Hoffmann found a pattern in performance that suggests students are disadvantaged by the need to scroll text, particularly students with lower computer skills.

Impact of Test Speededness. In a summary of 28 studies of 159 tests, Mead and Drasgow (1993) noted that comparability was most greatly impacted in speeded tests (tests which do not provide sufficient time for all examinees to finish). The studies they examined were of adult populations. One recent study found similar results for students in grades 4–12 taking a cognitive abilities test (Ito & Sykes, 2004). It should be noted that for most, if not all, state assessments, the students are allowed all the time they require so speededness should not be an issue.

Comparability for Different Student Subgroups

Concerns persist about whether there are subgroups of students who are disadvantaged because of lack of access, use, or familiarity with computers (Trotter, 2001). “Students will not take advantage of help options or use navigation guides if they require more personal processing energy than they can evoke” (Wissick and Gardner, 2000, p. 38).

The gap in access to technology is continuing to grow (Bolt & Crawford, 2000). The gap has widened considerably for computer ownership among racial minorities when compared with European-Americans. In the context of the overall racial digital divide, low-income European-American children are three times more likely to have Internet access than their African-American counterparts, and four times as likely as Latino children in the same socioeconomic category.

While equity is a critical concern, most studies do not focus on comparability for different subgroups of students, for a variety of logistical reasons.

Computer experience. For multiple-choice tests, the research to date suggests that differences in computer experience have little, if any, effect on test scores (Bridgeman, Bejar, & Friedman, 1999; Taylor, Jamieson, Eignor, & Kirsch, 1998; Russell, & Haney, 1997).

Findings from comparability studies show mixed results with computer-based constructed responses or writing assessments. Some studies suggested that students who had less experience with computers would do poorer on computer-administered tests (Russell, & Haney, 1997; Russell, 1999; Sandene et al, 2005). It is essential that prior training and practice be required. That omission may have contributed to the results of the Sandene study.

A study by O’Dwyer, Russell, Bebell & Tucker-Seeley (2005) suggests that after controlling for both prior achievement and socioeconomic status, students who reported greater frequency of technology use at school to edit papers were likely to have higher total

English/language arts test scores and higher writing scores. A recent study with NAEP assessment (Horkay, Bennett, Allen & Yan. 2006) found no significant mean score differences between paper and computer delivery. However, computer familiarity contributed significantly to predicting online writing test performance after controlling for paper writing skill. Other studies have found no evidence of such a disadvantage (Bennett, 2003; Higgins, Russell, & Hoffman, 2005).

Socio-Economic Status (SES): Sandene et al. (2005) found no significant difference in performance associated with parent's education level, a common proxy for SES. McCann[4] (2006) found a small SES interaction effect with proficiency. The interpretation of this interaction is, while SES by itself made no difference, low–achieving, low-SES students did about 1% less well on a computer-administered test while high-achieving, low-SES students were not affected by the mode of administration.

Ethnicity and Gender: Gallagher, Bridgeman, and Cahalan (2000) examined data from several national testing programs to determine whether the change from P&P to eTesting influences group differences in performance. Performance by gender, racial/ethnic, and language groups on the Graduation Record Examination (GRE) General Test, the Graduate Management Admissions Test (GMAT), the SAT I: Reasoning (SAT) test, the Praxis: Professional Assessment for Beginning Teachers (Praxis), and the Test of English as a Foreign Language (TOEFL) was analyzed. This study concluded the change is too small to pose a disadvantage to any of these subgroups.

However, some consistent patterns were found for some racial/ethnic and gender groups. African-American examinees and, to a lesser degree, Hispanic examinees appear to benefit from the eTesting format. However, for some tests, the eTesting version negatively affected female examinees. Ewing, Wiley, & Gillie (2003) found computer-based and paper-and-pencil math tests had the same factor structure for African-American, Asian, and Hispanic students. For each ethnic group, the same pattern of sub-scores emerged for a given total math score.

Ewing, et al. (2003) did find differences for English composition. For these scores, the pattern of sub-scores tended to vary for African-American, Asian, and Hispanic students. Neither Sandene et al. (2005) nor Nichols and Kirkpatrick (2005) found any differences in administration mode comparability among various demographic subgroups. Sim & Horton (2005), Sandene et al. (2005) and McCann (2006) failed to find any effect associated with gender.

Student Preferences: After students took computerized tests, some studies ask participants whether they would prefer to take future tests on computer or paper. In an evaluation of testing experience, students overwhelmingly preferred computer tests to paper tests (Brown & Augustine, 2001). The majority of students have indicated they would prefer to test on computer (Bridgeman, Lennon, & Jackenthal, 2001; Higgins, Russell, & Hoffman, 2005; Glassnapp, Poggio, Poggio, & Yang, 2005; Ito & Sykes, 2004; Johnson & Green, 2004; O'Malley et al., 2005; Pearson Educational Measurement, 2006; Sim & Horton, 2005; Wang, Young, & Brooks, 2004).

A survey of school staff, administered after a pilot study of the feasibility of administering the MCAS with Measured Progress iTest System for the Massachusetts Grade 7 students in Writing and Grade 10 students in Biology, concluded that students were very comfortable or somewhat comfortable using the iTest system. The majority of students who participated in the study adapted to the iTest delivery system quickly.

---

[4] McCann's study is based on Australian school children.

Students should have substantial time to practice and become familiar with the system in order to ensure valid results. The effect of typing ability on test scores may also have to be addressed. As one respondent indicated, "We need to figure out how to teach and develop keyboarding skills for all students. There is currently no keyboarding taught anywhere in our curriculum" (Massachusetts Department of Education, 2006).

Most students, regardless of demographics or ability, believed that the computer version was easier, faster, and more fun. Students also responded that using a computer helped concentration by presenting only one question at a time. A study at the Boston College Center for the Study of Testing, Evaluation, and Assessment found, "Students who are accustomed to writing on computers tend to do better on computerized tests than on paper exams. Conversely, students who do not use computers often to write tend to do better when they complete their tests on paper" (Trotter, 2001, p. 3).

Special Needs Students. A survey on computer use by students with disabilities in Germany (Ommerborn & Schuemer, 2001) reported more advantages than disadvantages to computer administration. Brown-Chidsey and Boscardin (1999) interviewed students with learning disabilities and found that the computer helped them with limitations that often interfered with the completion of their work. The research concluded, "Students' beliefs about computers are likely to shape the extent to which instructional technology enhances their achievement" (Brown-Chidsey, Boscardin, & Sireci, 1999, p. 4).

Summary

It is possible and appropriate to draw some generalizations from the published research.

- There appears to be a small but persistent advantage for students taking a multiple-choice test on computer.
- If scores from different modes are not interchangeable, it may be possible to make them comparable by simply shifting the eTesting scores by the effect size. However, more elaborate approaches such as multi-trait, multi-method techniques may prove more appropriate and effective.
- There are very few studies looking at direct measures of writing and those that do present inconsistent results. Of the three large studies of this issue, Kingston (2004) reported a moderate advantage for students taking a direct writing measure on computer; Sandene et al. (2005) reported no difference; Pearson Educational Measurement (2006) reported an advantage to students taking the test on paper.
- Electronic testing seems to present no disadvantage to students based on their gender, race, ethnicity, or socio-economic status. Additional, larger studies may be needed.
- The evidence regarding the impact of computer familiarity on performance on a electronic test is inconsistent. It may be minimized by giving students sufficient experience and realistic practice tests with the computer administration system before the test is administered.
- Teaching all students keyboarding skills may need to be included in the curriculum. Component 1 discusses an emerging trend in several states to develop technology standards for the curriculum.

- eTesting may disadvantage students when P&P testing-taking strategies no longer apply (Murphy, Long, Holleran, & Esterly, 2000; O'Malley, et al., 2005). Issues include necessity for scrolling, inability to highlight or mark out text, and differences in ancillary materials (e.g., calculators, rulers, graph paper).
- The inability to return to earlier items was a frequent complaint of students under early eTesting. Systems that permit item review and answer changes received much higher levels of student satisfaction.

It is perhaps trite but axiomatic that, for any type of testing with any mode of delivery, the assessment must match the instruction. When instruction is via a computer, students perform better when tested via a computer; when a computer is not integral to instruction, eTesting typically results in lower scores. This may compel the implementation of content standards for required technology-related curriculum.

COMPONENT 2 – REFERENCES AND LINKS

American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME) (1999). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

American Psychological Association (1986). *Guidelines for Computer-Based Tests and Interpretations*, Washington, DC: American Psychological Association.

Association of Test Publishers (2002). *Computer Based Testing Guidelines*, York, PA: Association of Test Publishers.

Bennett, R.E. (2003). *Online assessment and the comparability of score meaning.* Princeton, NJ: Educational Testing Service.

Bennett, R.E. (2002a). Inexorable and inevitable: The continuing story of technology and assessment. *Journal of Technology, Learning, and Assessment, 1*(1). Available from http://www.jtla.org.

Bolt, D., & Crawford, R. (2000). *Digital divide: Computers and our children's future*. New York: TV Books.

Bridgeman, B., Bejar, I. I., & Friedman, D. (1999). Fairness issues in a computer-based architectural licensure examination. *Computers in Human Behavior*, 15, 419–440.

Bridgeman, B., Lennon, M.L., & Jackenthal, A. (2001). *Effects of screen size, screen resolution, and display rate on computer-based test performance* (ETS RR-01-23). Princeton, NJ: Educational Testing Service.

Brown-Chidsey, R., & Boscardin, M. L. (1999). *Computers as accessibility tools for students with and without learning disabilities.* Amherst, MA: University of Massachusetts.

Brown-Chidsey, R., Boscardin, M. L., & Sireci, S. G. (1999). Computers attitudes and opinions of students with and without disabilities. Amherst, MA: University of Massachusetts.

Brown, P. J., & Augustine, A. (2001). *Screen reading software as an assessment accommodation: Implications for instruction and student performance*. Paper presented at the American Education Research Association Annual Meeting, Seattle, WA, April, 2001.

Cerillo, T.L. & Davis, J.A. (2004). *Comparison of paper-based and computer based administrations of high-stakes, high-school graduation tests.* Paper presented at the Annual Meeting of the American Education Research Association, San Diego, CA.

Choi, S.W. & Tinkler, T. (2002). *Evaluating comparability of paper and computer-based assessment in a K–12 setting.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Donovan, M. A., Drasgow, F., Probst, T. M. (2000). Does computerizing paper-and-pencil job attitude scale make a difference? New IRT analyses offer insight. *Journal of applied Psychology*, 85(2), 305-313.

EdWeek (2003a). *Legal Twists, Digital Turns, 2003 Editorial Projects in Education,* 22(35), pp. 11–16, 16. Retrieved on 02/10/06 from http://counts.edweek.org/sreports/tc03/article.cfm?slug=35impact.h22.

EdWeek (2003b). *Pencils Down: Technology's Answer to Testing*. Retrieved on 02/10/06 from http://counts.edweek.org/sreports/tc03/article.cfm?slug=35exec.h22.

EdWeek (2003). *Tech's Answer to Testing, 2003 Editorial Projects in Education,* 22(35), pp. 8, 16. Retrieved 02/10/06 from http://counts.edweek.org/sreports/tc03/article.cfm?slug=35.h22.

EdWeek (2005). *Testing Computerized Exams. Education Week on the Web. ,* Retrieved on 02/20/2007 from: http://www.edweek.org/chat/transcript_12_14_2005.html?levelld=2200&levelld=1000.

Ewing, M., Wiley, A. & Gillie, J.M. (2003). *Moving from paper-and-pencil administration to computer-based testing: An investigation of construct equivalence and subgroup differences.* Paper presented at the 2003 NCME Annual Meeting, April 2003, Chicago, IL

Gallagher, A., Bridgeman, B., and Cahalan, C. (2000). *The Effect of Computer-Based Tests on Racial/Ethnic, Gender, and Language Groups (ETS Research Rep. No. 00-8).* Reviewed 3/10/07 from http://www.ets.org/Media/Research/pdf/RR-00-08-Gallagher.pdf.

Glassnapp, D.R., Poggio, J., Poggio, A., & Yang, X. (2005). *Student attitudes and perceptions regarding computerized testing and the relationship to performance in large scale assessment programs.* Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal.

Goldberg, A., Russell, M., & Cook, A. (2003). The effect of computers on student writing: A metaanalysis of studies from 1992 to 2002. *Journal of Technology, Learning, and Assessment, 2*(1). Available from http://www.jtla.org.

Higgins, J., Russell, M., & Hoffmann, T. (2005). Examining the effect of computer-based passage presentation on reading test performance. *Journal of Technology, Learning, and Assessment, 3*(4). Retrieved on 3/30/06 from http://www.jtla.org.

Horkay, N., Bennett, R E., Allen, N., Kaplan, B., & Yan, F. (2006). 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(2), Retrieved 1/20/07 from http://www.jtla.org.

Ito, K. & Sykes, R.C. (2004). *Comparability of scores from norm-referenced paper-and-pencil and web-based linear tests for grades 4-12.* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Johnson, M. & Green, S. (2004). *On-line assessment: the impact of mode on students' strategies, perceptions, and behaviors*. Paper presented at the annual meeting of the British Educational Research Association, Manchester, Great Britain.

Kim, D. & Huynh, H. (2006). *A comparison of student performance between paper-and-pencil and computer-based testing in four content areas*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Kingston, N.M. (2004). *The comparability of scores of a direct writing measure administered on computer versus administered on paper.* Unpublished report presented to the Maine Department of education.

Kingston, N.M. (2002). *Comparability of scores from computer- and paper-based administrations for students in grades K-8*. 32nd annual Large-Scale Assessment Conference of the Council of Chief State School Officers, Palm Desert, CA.

Kolen, M. J. (1999-2000). Threats to score comparability with applications to performance assessments and computerized adaptive tests. *Educational Assessment*, 6, 73–96.

Makiney, J. D., Rosen, C., Davis, B. W., Tinios, K. & Young, P. (2003). *Examining the measurement equivalence of paper and computerized job analyses scales.* Paper presented at the 18th Annual Conference of the Society for Industrial and Organizational Psychology, Orlando, FL.

Massachusetts Department of Education (2006). *Report and Analysis of the MCAS Online Pilot Feasibility Study Using the Measured Progress iTest System*.

McCann, R. (2006). The equivalence of online and traditional testing for different subpopulations and item types. *British journal of Educational Technology*, 37(1), 79–81.

Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin, 114*(3), 449–458.

Merten, T. (1996). A comparison of computerized and conventional administration of the German versions of the Eysenck Personality Questionnaire and the Carroll Rating Scale for Depression. *Personality and Individual Differences*, 20, 281–291.

Murphy, P.K., Long, J., Holleran, T., & Esterly, E. (2000). *Persuasion online or on paper: A new take on an old issue*. Paper presented at the annual meeting of the American Psychological Association, Washington, DC.

New England Compact (February, 2005) *Computers Improve Students' Writing Performance on Tests.* Education Development Center: Newton, MA. Retrieved May 10, 2007 from http://www.necompact.org/NECompact%20Summaries_0205.pdf

Nichols, L.M. (1996). Pencil and paper versus word processing: a comparative study of creative writing in elementary school. *Journal of Research on Computing in Education, 29,* 59–66.

Nichols, P. & Kirkpatrick, R. (2005). *Comparability of the computer-administered tests with existing paper-and-pencil tests in reading and mathematics tests*. Paper presented at the AERA Division D Graduate Student Seminar, Montreal, Canada.

O'Dwyer, L., Russell, M., Bebell, D., & Tucker-Seeley, K. R. (2005). Examining the reliationship between home and school computer use and students' English Language Arts Test Scores. *Journal of Technology, Learning, and Assessment*, 3(3), 4–45.

O'Malley, K. J., Kirkpatrick, R., Sherwood, W., Burdick, H. J., Hsieh, M.C., Sanford, E.E. (2005). *Comparability of a Paper Based and Computer Based Reading Test in Early Elementary Grades*. Paper presented at the AERA Division D Graduate Student Seminar, Montreal, Canada.

Ommerborn, R., & Schuemer, R. (2001). *Using computers in distance study: Results of a survey amongst disabled distance students.* FernUniversitat-Gesamthochschule in Hagen. Retrieved 2/1/2007, from: http://www.fernuni-hagen.de/ZIFF

Paek, P. (2005). *Recent trends in comparability studies* (PEM Research Report 05-05). Available from http://www.pearsonedmeasurement.com/downloads/research/RR_05_05.pdf.

Pearson Educational Measurement. (2002). *Final report on the comparability of computer-delivered and paper tests for Algebra I, Earth Science and English.* Austin, TX: Author.

Pearson Educational Measurement. (2003). *Virginia standards of learning web-based assessments comparability study report – Spring 2002 administration: Online & paper tests*. Austin, TX: Author.

Pearson Educational Measurement (2006). *Comparability analysis of online and paper-and-pencil MEAP.* Iowa City, IA: Author

Pinsoneault, T. B. (1996). Equivalency of computer-assisted paper-and-pencil administered version of the Minnesota Multiphasic personality Inventory-2. *Computers in Human Behavior*, 12, 291–300.

Poggio, J., Glasnapp, D.R., Yang, X. & Poggio, A.J. (2005b). A comparative evaluation of score results from computerized and paper & pencil mathematics testing in a large scale state assessment program. *The Journal of Technology, Learning, and Assessment, 3*(6), 1–30.

Pommerich, M. (2004). Developing computerized versions of paper tests: Mode effects for passage-based tests. *The Journal of Technology, Learning, and Assessment, 2*(6), 1–44.

Pomplun, M., Frey, S. & Becker, D.F. (2002). The score equivalence of paper and computerized versions of a speeded test of reading comprehension. *Educational and Psychological Measurement, 62*(2), 337–354.

Power, D. E., & Potenza, M. T. (1996). *Comparability of testing using laptop and desktop computers (RR-96-15).* Princeton, NJ: Educational Testing Service.

Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: A comparison of confirmatory factor analysis and item response theory. *Journal of Applied Psychology*, 87 (3), 517–529.

Russell, M. (1999). Testing on computers: A follow-up study comparing performance on computer and on paper. *Education Policy Analysis Archives, 7*(20). Retrieved 2/20/2007, from http://epaa.asu.edu/epaa/v7n20

Russell, M. (2002). It's Time to Upgrade: Tests and administration procedures for the new millennium. *Essays in Education*, 1(1). Retrieved February, 2007: http://www.columbiacollegesc.edu/essays/vol.%201%20spring%202002/time_to_upgrade_revised.pdf.

Russell, M. & Hoffmann, T. (2005). Examining the Effect of Computer-Based Passage Presentation on Reading Test Performance. *Journal of Technology, Learning, and Assessment*, 3(4). Retrieved 3/25/07, from http://www.jtla.org

Russell, M. & Plati, T. (2001). Effects of computer versus paper administration of a state-mandated writing assessment. *Teachers College*. Retrieved February, 2005, from http://www.tcrecord.org

Russell M. & Haney. W. (2000a). Bridging the gap between testing and technology in schools. *Education Policy Analysis Archives, 8*(19). Retrieved February 1, 2005, from http://epaa.asu.edu/epaa/v8n19.html

Russell, M. & Haney, W. (1997). Testing writing on computers: An experiment comparing student performance on tests conducted via computer and via paper . *Educational Policy Analysis Archives, 5*(3). Retrieved February 1, 2005, from http://epaa.asu.edu/epaa/v8n19.html

Sandene, B., Horkay, N., Bennett, R., Allen, N., Kaplan, B., and Oranje, A. (2005). Online assessment in mathematics and writing: *Reports from NAEP Technology-Based Assessment Project, Research and Development Series* (NCES 2005-457). U.S. Department of Education, National Center for Educational Statistics. Washington, DC: U.S. Government Printing Office. Retrieved on 3/25/07 from http://nces.ed.gov/nationsreportcard/pdf/studies/2005457_1.pdf.

Schaeffer, G. A., Bridgeman, B., Golub-Smith, M. L., Lewis, C., Potenza, M. T., & Steffen, M. (1998). *Comparability of paper-and-pencil and computer adaptive test scores on the GRE General Test (RR-98-38)*. Princeton, NJ: Educational Testing Service.

Sim, G. & Horton, M. (2005) *Performance and Attitude of Children in Computer Based Versus Paper Based Testing*. World Conference on Educational Multimedia, Hypermedia and Telecommunications, Montreal.

Stocking, M. (1997) Revising item responses in computerized adaptive tests: a comparison of three models. *Applied Psychological Measurement. 21*, 129–142.

Taylor, C., Jamieson, J., Eignor, D., & Kirsch, I. (1998). *The relationship between computer familiarity and performance on computer-based TOEFL test tasks (TOEFL-RR-61).* Princeton, NJ: Educational Testing Service.

Trotter, A. (2001). Testing computerized exams. *Education Week*, 20 (37) 30–35. Retrieved March, 2002, from: http://www.edweek.org/ew/westory.cfm?slug=37online.h20

U.S. Department of Commerce. (2002). *A nation online: How Americans are expanding their use of the Internet*. Washington, DC: Author. Retrieved April 19, 2002, from http://www.ntia.doc.gov/ntiahome/dn/nationonline_020502.htm

U.S. Department of Education, Office of Planning, Evaluation and Policy Department U.S. Department of Education (2007). *State Strategies and Practices for Educational Technology: Volume I—Examining the Enhancing Education Through Technology Program*. Washington, DC: Author.

Vispoel, W. P., Clough, S. J., Bleiler, T., Hendrickson, A. B., & Ihrig, D. (2002) Can examinees use judgments of item difficulty to improve proficiency estimates on computerized adaptive vocabulary tests? *Journal of Educational Measurement, 39,* 311–330.

Vispoel, W., Hendrickson, A. B., & Bleiler, T. (2000). Limiting answer review and change on computerized adaptive vocabulary test: Psychometric and attitudinal results, *Journal of Educational Measurement,* 37, 21–38.

Wang, S. (2004). *Online or paper: Does delivery affect results? Administration mode comparability study for Stanford diagnostic Reading and Mathematics tests.* San Antonio, Texas: Harcourt.

Wang, S., Young, M., and Brooks, T. (2004). *Administration mode comparability study for Stanford Diagnostic Reading and Mathematics Tests*. San Antonio, Texas: Harcourt.

Way, D. D., Davis, L. L., & Fizpatrick, S. (2006). *Practical Questions in Introducing Computerized Adaptive Testing for K-12 Assessments*. Retrieved 3/26/07 from http://www.pearsonedmeasurement.com/research/research.htm#

Williams, S. (n.d.) *Online Testing*. Retrieved 3/26/2007 from www.indiana-etc.org/pdfs%5COnline-Testing.pdf

Wissick, C. A., & Gardner, J. E. (2000). Multimedia or not to multimedia? That is the question for students with learning disabilities. *Teaching Exceptional Children*, 32 (4), 34–43.

Yang, X., Poggio, A. Glasnapp, D., & Poggio, J. (2007) *Score comparability between computerized and paper-and-pencil linear tests.* Unpublished paper.

Zandviliet, D., & Farragher, P. (1997). A comparison of computer-administered and written tests. *Journal of Research on Computing in Education*, 29(4), 423–438.

COMPONENT 3 – A DETERMINATION AS TO WHETHER THE CONVERSION OF THE STATE ASSESSMENT PROGRAM TO A COMPUTER-BASED OR COMPUTER-ADAPTIVE FORMAT WILL SATISFY THE FEDERAL NO CHILD LEFT BEHIND REQUIREMENTS

Current NCLB Requirements

No requirement in the current NCLB legislation specifically precludes computer-based testing or computer-adaptive testing, as long as the computerized testing meets all NCLB requirements that pertain to all tests, whether they are administered via computers or paper/pencil (e.g., tests are aligned with state standards and on-grade-level). However, there are NCLB requirements that are particularly relevant to states transitioning to computerized testing. These include requirements for comparability among modes of administration (i.e., no construct-irrelevant variance), assurances of score comparability for Adequate Yearly Progress (AYP) reporting consistency via equating or otherwise, and on-grade-level assessment items for AYP reporting purposes.

Comparability

NCLB requires documented evidence of the comparability of the computer-based or computer-adaptive test administration with previous or concurrent paper-and-pencil test administrations. In this sense, the paper and pencil version of the test serves as the gold standard, more so as an existing condition or historical fact than anything else. Establishing comparability would be important even if the state were to transition entirely to computerized testing in a single year, since the previous score data, and associated information, reported both within the state and to the U.S. Department of Education, would have been from paper/pencil administered tests. This requirement can be found in the USDOE Standards and Assessment Peer

Review Guidance under Peer Review Critical Element 4.4(b) which states *"If the State administers both an online and paper and pencil test, has the State documented the comparability of the electronic and paper forms of the test?"* (USDOE, 2004). Meeting these comparability requirements is not insurmountable, as many states have done so successfully through careful planning and using psychometrically sound methodologies. Please see Component 2 for a more detailed discussion of comparability studies and the potential impact on South Carolina's assessment program, should the state decide to make the transition to computerized testing.

Construct-Irrelevant Variance (Testing Mode Effect)

Secondly, a transition to computerized testing must ensure that the test administration method does not interfere with a student's test performance (i.e., construct-irrelevant variance). This is addressed in the USDOE Standards and Assessment Peer Review Guidance under Peer Review Critical Element 4.4(g) which states, *"Has the State ascertained whether the assessment produces intended and unintended consequences?"* (USDOE, 2004).

Concern of a possible mode-of-administration effect attributable to the computerized administration of a statewide assessment stems from a number of probable circumstances. First, it is possible that the interaction between the test taker and the computer may interfere with test performance. It is important in any test to ensure that the response mode itself does not interfere with what is being measured. It is also important to ensure that the mode of administration itself (i.e., the computer) does not introduce unintended and unforeseen variables into the testing situation. For example, results from comparability studies should help to address the question of whether students who have had less computer experience are impacted differently by the electronic test administration.

The intent of a computerized test is to measure students in terms of educational achievement and not to assess computer expertise. Associated with this concern are issues regarding how the computerization of a test may interfere with learned test-taking strategies. This must be weighed against the extent to which newer test administration software has integrated features to mitigate this effect. Finally, there may be issues related to whether item parameter estimates from one mode of administration are directly transferable to another. Ultimately, however, the main concern is the degree of equivalence between test scores from computerized and paper-and-pencil versions of the same test.

With careful planning and implementation, these concerns can be addressed to satisfy NCLB requirements.

Specific Issues Related to Computer-Adaptive Testing and NCLB

Lastly, the USDOE has determined that all test items used for AYP determinations must be on-grade-level and measure a component of the state's content standards for that grade. In the traditional CAT environment, the computer algorithm that distributes the items is used to determine a student's performance level on a particular skill, irrespective of the assigned grade-level for that skill. That is, the CAT algorithm selects and administers the item closest to the student's estimated level of achievement regardless of the grade nominally associated with the item. This topic is addressed more fully in the section devoted to CAT at the end of the report.

Two states currently use some variation of computer-adaptive testing and still meet the on-grade-level requirements of NCLB: Idaho and Oregon.

IDAHO

The Idaho Standards Achievement Test (ISAT) was originally designed as a pure CAT program and was initially rejected by the U.S. Department of Education as not meeting the requirements of NCLB (USDOE, December 9, 2005). The ISAT was initially rejected for three main reasons:

- The ISAT was not aligned to the Idaho content standards. A third-party review of the ISAT test items and Idaho content standards revealed that the match between the test items included on the ISAT and the state's content standards was not sufficiently strong to deem it "aligned," which is one of the critical requirements for NCLB.
- Since the ISAT was solely a CAT, it contained items that were not on-grade-level. Rather, the test was designed to measure the student at his/her performance level, irrespective of the student's grade level. Such a test design did not meet the NCLB requirement for determining AYP and/or grade-level proficiencies.
- The ISAT was initially administered over a six-week period, which was deemed too long in that students testing later in the window could have benefited from as many as six additional weeks of instruction than their peers who tested early in the window.

To bring the ISAT into NCLB compliance, the state of Idaho took the following actions:

- Altered the content of the ISAT to include a sufficient number of items directly targeted at measuring the Idaho content standards.

- Required that Idaho students testing in the NCLB-required grades and subjects take a common set of on-grade-level items. These scores are currently used for NCLB reporting and AYP calculations. The state did retain the CAT elements of the previous ISAT by including them in a separate test section. The CAT items continue to be selected by the computer program according to the student's performance on the common core items. However, the CAT scores are reported separately. While it is not a part of NCLB requirements, a fall/winter CAT is administered using the same basic design as the spring CAT section.
- Reduced the testing window by one week during the first year of administration of the new ISAT.

Upon completion of the above revisions to the assessment system, the state's assessment program was granted full approval by the USDOE in 2006 (USDOE, November 16, 2006).

Oregon

The state's Technology Enhanced Student Assessment (TESA) is a modified CAT, with all items within the system being on-grade-level and aligned with state content standards (unlike Idaho, which has an adaptive component added to a core set of grade-level items taken by all students). These assessments are available to all students in participating schools in grades 3–8 and 10, and include reading/literature, mathematics, science, and social science; the performance writing tests are administered solely via paper/pencil.

As of June 22, 2006, the USDOE had not given full approval to the Oregon assessment system, citing "concerns with the alignment of standards to grade-level content standards and the technical quality, including validity, reliability, and comparability of assessments in varying formats." (USDOE, June 22, 2006) However, on April 25, 2007, Oregon published/updated its

2005-2006 Technical Report, Volume 2, Test Development and *2006-2007 Technical Report, Volume 4, Reliability and Validity*. These publications include information and comparability studies that address the USDOE concerns, and in an e-mail correspondence to DRC, dated June 4, 2007, the Oregon Department of Education indicated it is anticipating full approval by the USDOE.

The following description of TESA is drawn from the *2005-2006 Technical Report, Volume 2, Test Development* (Oregon, Vol. 2, 2007).

DESCRIPTION OF TESA

The TESA became computer-based in 1999 and adaptive in 2001. In 2005-2006, 90% of Oregon students tested via computer. Students are provided with up to three opportunities (from September to May) to take the standard grade-level test using TESA, and if a student tests more than once, the student's highest score is retained. (Note: In spring of 2007, issues with Oregon's computerized test delivery system prompted the Department of Education to revert to all paper/pencil administrations for this spring. See Component 1 for a more detailed discussion of this topic.)

"As with Oregon's paper/pencil test forms, each Score Reporting Category (SCR) for every content-area test "is represented by a specific percentage of items and has items from a range of difficulty levels [that] are approximately equivalent across SRCs. . . . Pools contain enough items in the SRCs to allow for representative selection within each category, allowing for items to be adaptively selected while maintaining the SRC weighting [test blueprint] for each test. . . . The accuracy of the student responses to items determines the next block of items and passage the student will see. Thus, each student is presented with a set of items that most accurately aligns with his or her ability level." (Oregon, Vol. 2, 2007, p. 23.)

“…[E]ach grade has three equivalent adaptive pool of items, one for each retake opportunity. The adaptive test pools are designed so that students can meet the standards regardless of the item bank from which items are drawn. . . . [E]ach pool contains a percentage of items, typically approximately 80%-90%, that have been previously used operationally and are psychometrically sound. The criteria for these items are the same as those for selecting anchor items for the paper-and-pencil test forms.” (Oregon, Vol. 2, 2007, p. 23.)

“The TESA item banks each contain approximately 300 items per grade and subject (with the average test being 50-60 items long), a sufficient number . . . to ensure that students are administered more or less difficult tests and are provided items representing the breadth and depth identified in the test specifications and content standards.” (Oregon, Vol. 2, 2007, p. 25.)

The Future of NCLB

The No Child Left Behind Act (NCLB) is slated for reauthorization in fiscal year 2008; thus, the reauthorization will take place subsequent to the release of this study. As expected, many entities are weighing in on suggested revisions to NCLB. These organizations include, among others, the American Federation of Teachers, the 50 State Teachers of the Year, the Commission on the No Child Left Behind Act (Aspen Institute), the Council of Great City Schools, Education Trust, the Forum on Educational Assessment, the National Association of Secondary School Principals, the National Education Association, and the National School Boards Association. The Council of Chief State School Officers, the National Association of State Boards of Education, and the National Governors’ Association recently released a joint statement on the reauthorization of NCLB. Additionally, legislators from various states are espousing positions ranging from making state participation in NCLB voluntary, to providing more flexibility in grades to be tested and models to be used for calculating Adequate Yearly

Progress, to calling for a set of national standards and tests with directs link to the National Assessment of Educational Progress.

Of all the suggested amendments/policy suggestions related to NCLB reauthorization, a few address issues related to computer-adaptive testing.

Testimony [before Congress] concerning the measurement of student achievement growth, Allan Olson, Co-Founder and [former] Chief Academic Officer, Northwest Evaluation Association (Olson, 2007):

> Can we measure achievement growth of individual students? It is clear that two components are needed to measure the achievement growth of individual students. The first requirement is the ability to measure students accurately to gain a deep understanding of where their learning is. Current tests provide little information about students who are high performers and are well beyond their grade level or low performers who are well behind grade level. To be able to measure achievement for these students requires a measurement scale that goes beyond grade-level testing and identifies what students know across the many strands of knowledge that a student needs to know to be identified as proficient.

Dr. Charlene Rivera, Executive Director, Center for Equity and Excellence in Education, The George Washington University, testimony before the Committee on Health, Education, Labor, and Pensions (HELP) Meeting on Reauthorization of NCLB (Rivera, 2007):

> The administration of English Language Arts (ELA) assessments to ELLs, while appropriate, is a source of controversy. Options for allaying the controversy include allowing states to create alternate assessments of ELA for ELLs, linked to grade level content standards, and computer adaptive versions of ELA tests. In the case of the latter, the computer quickly determines the student's overall level of mastery, and tailors the test questions to that level. The law could encourage states to explore such options.

Joint Statement of the National Conference of State Legislatures and the American Association of School Administrators on ESEA Reauthorization (NCSL & AASA, 2007):

> NCSL and AASA believe that Title I should support flexibility for states and school districts in using a variety of standards-based assessment and accountability systems that measure the academic progress of individual students, including value-added models, benchmarking models, computer-adaptive assessments and instructionally sensitive assessments.

Despite the statements above, there is little indication that reauthorization will include any increased flexibility for high-stakes, content-area testing via CAT, as standards-aligned, grade-level assessments are at the core of NCLB. A review of Secretary Spellings' April 23, 2007, letter outlining USDOE policy priorities for reauthorization indicates that an expansion of the growth model pilot program and new funding priorities are more likely changes (Spellings, 2007).

Post-Equating versus Pre-Equating

It may also be necessary to reconsider the method by which South Carolina equates test forms. Many states, including South Carolina, rely on a so-called post-equating model. This approach uses field test results to build approximately parallel forms to balance content and prevent drastic shifts in test difficulty from form to form (year to year, typically). However, the final equating to determine the exact cut-score for the current year's form is done after the operational administration, letting the data from the most recent administration determine the exact difficulties of the items and the mean shift needed to equate to the previous forms.

The advantages of post-equating include using the current context of the items and guarding against unforeseen disturbances (e.g., an item being made public). Events inside or outside the classroom may have altered the relative difficulties of items. Nuisance factors like item sequence in the test and differences between field test and operational test motivation can also have an effect. Post-equating bases the final scaling within the context that the students actually experienced.

Alternatively, some states use a pre-equating model. This approach relies heavily on the field test results. All item parameters are estimated from the field test and are considered fixed for the life of the item. In a fixed-form environment, CBT or not, this allows all forms to be created and scaled prior to any operational administration. When the item parameter estimates are established before testing, scores can be reported immediately after testing. Instant reporting is a major attraction of pre-equating, although it loses some its impact if reporting must wait for the completion of hand scoring of constructed-response items.

Some test industry experts recommend pre-equating precisely because it does not adapt to the current context. That is, the original scaling of the items reflected the situation at the beginning of the intervention. Rescaling could remove the effect of improved instruction if it affects items to differing degrees. For example, a major innovation in teaching of fractions might make items dealing with fractions appear easier rather than reflect an improvement in student performance.

CAT depends on pre-calibrated items to guide the selection of the best items to present next. While there are strategies to calibrate *on the fly*, most high-stakes item banks used in CAT applications rely on item parameter estimates being established prior to operational testing. Working with established parameter estimates ensures that any form drawn from the bank will be pre-equated to every other possible form. Once operational, replacement items can be continuously tried out and cycled into the bank as a routine part of testing.

Pre-equating places a heavy burden on the field-test process:

- The item parameter estimates must be established and in the final metric prior to testing.
- Calibration samples need to be sufficiently large to provide stable, reliable estimates.
- The context for the student must be very similar to the operational context, including time of year, time of day, test format, and student motivation.

Diagnostic reports could conceivably be provided immediately after testing, even without pre-equating, if they did not include final scale scores, performance level classifications, or hand-scored CR. Final accountability results could thus wait for completion of the post-equating process.

Summary

There is nothing in the current NCLB legislation and regulations/guidance that prohibits the use of either CBT or CAT testing, as long as certain requirements are met (e.g., comparability among modes; grade-level items for AYP purposes; and demonstrated, psychometrically sound equating practices). The trade-offs involved in any shift to CBT/CAT testing are addressed in other sections of this report.

Any transition to CBT or CAT by the state of South Carolina will require a review of the state's current item bank. Considerations must include an evaluation of each item to ascertain which ones can be transitioned appropriately to a computer environment. Additionally, should South Carolina wish to pursue computer-adaptive tests (whether modeled on Idaho's or Oregon's approach), a full review of the current South Carolina item bank will be necessary, with focus on the robustness of items, content coverage, and level of difficulty.

Due to the proprietary nature of South Carolina's current item banks for PACT, HSAP, and end-of-course tests, DRC was not able to conduct full computer-adaptive test simulations using the state's on-grade item pool within each grade-level content area in order to determine whether there are sufficient items to support a computer-adaptive test format that would meet federal requirements. This question was also not pursued because the final recommendation of this study is not to implement CAT at this time. If the SDE decides to shift to CAT in the future, the viability of the item bank at that time for CAT should be investigated.

Full reporting immediately after testing with either CBT or CAT will require pre-equating, which is a change in South Carolina's approach to equating. CAT relies on pre-equated items to guide the selection of the next item. A post-equating approach would still be possible with CBT if the early reports were diagnostic, preliminary, and did not depend on the final scaling.

COMPONENT 3 – REFERENCES AND LINKS

Joint Statement of the National Conference of State Legislatures and the American Association of School Administrators on ESEA Reauthorization (NCSL & AASA, 2007). Available at: http://www.ncsl.org/statefed/NCSLAASAJointStatement.htm

Olson, A. (2007) "Testimony [before Congress] concerning the measurement of student achievement growth," Allan Olson, Co-Founder and Chief Academic Officer, Northwest Evaluation Association (March 21, 2007). Available at: http://edlabor.house.gov/testimony/032107AllanOlsontestimony.pdf

Oregon State Department of Education (2007) *2005-2006 Technical Report, Volume 2, Test Development.* Available at: http://www.ode.state.or.us/teachlearn/testing/manuals/2007/asmttechmanualvol2_testdevelopment.pdf

Oregon State Department of Education (2007) *2006-2007 Technical Report, Volume 4, Reliability and Validity.* Available at: http://www.ode.state.or.us/teachlearn/testing/manuals/2007/asmttechmanualvol4_validity.pdf

Rivera, C. (2007). Testimony before the Committee on Health, Education, Labor, and Pensions (HELP) Meeting on Reauthorization of NCLB, April 4, 2007. Available at: http://www.nekia.org/files/web_CEEE_RiveraSenate_NCLBStatement_4_10_07.doc

Spellings, M. (2007). NCLB reauthorization policy priorities. Letter to Chief State School Officers, April 23, 2007. Available at http://www.ed.gov/policy/elsec/guid/secletter070423.html.

United States Department of Education (2004) "Standards and Assessments Peer Review Guidance: Information and Examples for Meeting the Requirements of the No Child Left Behind Act of 2001." Available at http://www.ed.gov/policy/elsec/guid/saaprguidance.pdf

United States Department of Education (December 9, 2005), Policy Letter to Idaho. Available at: http://www.ed.gov/admins/lead/account/nclbfinalassess/id.html

United States Department of Education (November 16, 2006), Policy Letter to Idaho. Available at: http://www.ed.gov/admins/lead/account/nclbfinalassess/id3.html

United States Department of Education (June 22, 2006), Policy Letter to Oregon. Available at: http://www.ed.gov/admins/lead/account/nclbfinalassess/or2.html

COMPONENT 4 – RECOMMENDATIONS REGARDING SUBJECT AREA ASSESSMENTS TO BE COMPUTER-BASED OR COMPUTER ADAPTIVE TO INCLUDE A RECOMMENDATION REGARDING ORDER OF IMPLEMENTATION

The specifics of the situation in South Carolina will ultimately determine the best course of action for the phase-in of electronic testing (eTesting) for South Carolina. Any assessment will benefit from quality items and presentation, solid diagnostic capability, and shortened reporting schedules. To aid in the decision about the sequence of grades and content areas, three sources of information about the optimal sequence of content areas and grades have been considered to frame the recommendations for implementation:

- Literature Review of Comparability Studies,
- Questionnaire and Discussion Groups with SC stake holders, and
- Expert Panel Meeting.[1]

Comparability Studies and Content Areas

Assuming that the published comparability studies reflect the grades and content areas that were first implemented, the literature review should indicate this. However, most published studies have not dealt with state-mandated assessments and most do not indicate why those particular choices were made. Reading and mathematics were well represented in the research. This may reflect the importance of these areas in the states' curricula and assessments as well as

[1] The Expert Panel of State Departments of Education staff directly involved in CBT was convened by DRC specifically for the feasibility study. Florida, Georgia, Idaho, Indiana, Kentucky, North Carolina, Oklahoma, Virginia, and West Virginia were represented. Kansas was represented by the University of Kansas group that assists the SDE with all phases of the assessment. The information sought from this group was what things should be considered when deciding if to proceed with CBT and, if SC decides to proceed, what things are important to do and what things are important to avoid.

NCLB concerns and requirements, not necessarily, because they are the best choices for early phase-in for a large-scale assessment program.

Writing was the subject of a number of studies and is a special case because it relies almost exclusively on extended written responses (ER). While there were fewer studies dealing with this format than there were for multiple choice, and these reached different conclusions, there was a consistent finding that the students with more computer experience did better on computer-based direct writing assessments (e.g., see New England Compact, 2005). The important factors were level of computer proficiency and match with instruction. Students taught writing on computer performed better when tested on computers and the higher the student achievement, the more the gain.

Educator Conference Questionnaires

While the primary data source for information on the feasibility study is the Web survey that all districts were requested to complete, a more informal questionnaire was distributed at two conferences for front line educators in the state. These conferences were the South Carolina Educators for the Practical Use of Research (SCEPUR) Annual Conference and the South Carolina Middle School Association (MSA) Annual Conference. While the focus of the Web survey was to collect hard data on numbers of computers and Internet capability, the intent of the questionnaire was to obtain a general impression of the attitudes toward eTesting and to identify any unforeseen circumstances that might work against or for a successful implementation. Detailed questionnaire data and respondent comments are included on the attached CD.

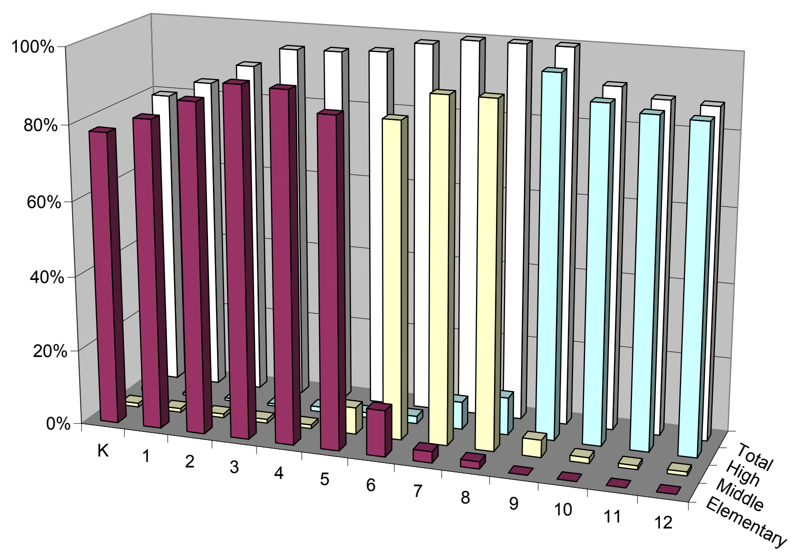
Questionnaires were collected from 123 school and district personnel and the question was explicitly asked, "Which grade and content area should be implemented first?" The majority of respondents indicated eTesting should begin with reading or math because these are the most basic and receive the most emphasis, both in instruction and assessment. This opinion may be influenced by the expectation that eTesting is synonymous with the rapid reporting of test results.

Social studies was also a popular choice. The comments suggested that the respondents felt it would be easier to assess than more graphical areas like science.

Writing received very few votes because teachers felt students lacked the experience and training to compose on computers. This is inconsistent with the findings of the comparability studies and the results of the Web survey of schools and districts. The survey responses indicated that students begin receiving training and using computers effectively as soon as they enter school.

There also was some wariness about the objectivity of scoring for computer-based essays. While the skepticism was greatest for computer scoring, there are also reservations about the objectivity of essay scores under any circumstances.

Figure 4.1: Computer-Based Instruction by Grade



In response to the question about the appropriate grade to begin eTesting, most respondents indicated it should begin in grade 3. Students by that point should have adequate experience and coordination to manage appropriate computerized tests.

Respondents from middle and high school indicated eTesting should begin with the lowest grade in the school, e.g., six or nine. Because the questionnaire instructions were to "respond for your school," this response can be interpreted to mean electronic testing is feasible at the lowest grade in the school and perhaps the grades in the lower schools as well.

The dominant theme in the comments was that eTesting should begin as early as possible to provide experience and to ensure a high level of comfort as students progress. However, there was an important counterpoint: eTesting should not begin too early; it should not begin before all students, regardless of the resources both in and out of school, have had sufficient exposure and training to ensure a fair and equitable assessment. Again, the Web survey indicated widespread computer usage and training from kindergarten on. Figure 4.1 shows that 78% of the kindergarteners and 96% of ninth graders represented by respondents to the Web Survey were receiving computer-based instruction.

Figure 4.2: Frequency of Computer-Based Writing Practice

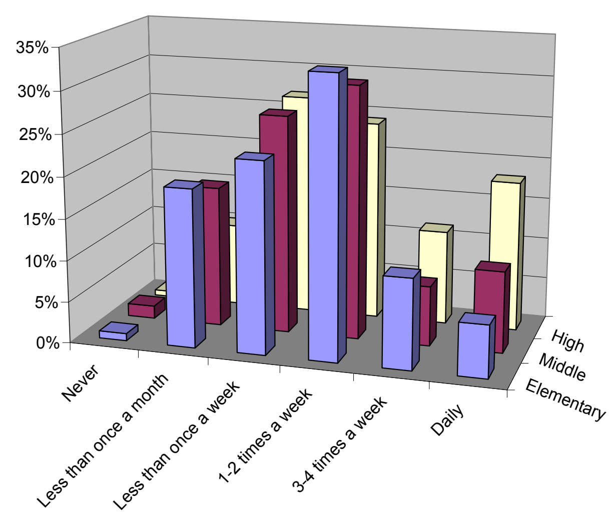


Figure 4.2 shows that almost all students receive some writing practice on computer; however, about 20% of elementary students receive such practice less than once a month. Across all grades, about 50% of students practice writing on computer at least once per week.

Whether this is adequate to provide the appropriate match with instruction or to provide sufficient computer proficiency to support computer-based writing assessment may require additional discussion.

Figure 4.3: Time Spent on Computers[2]

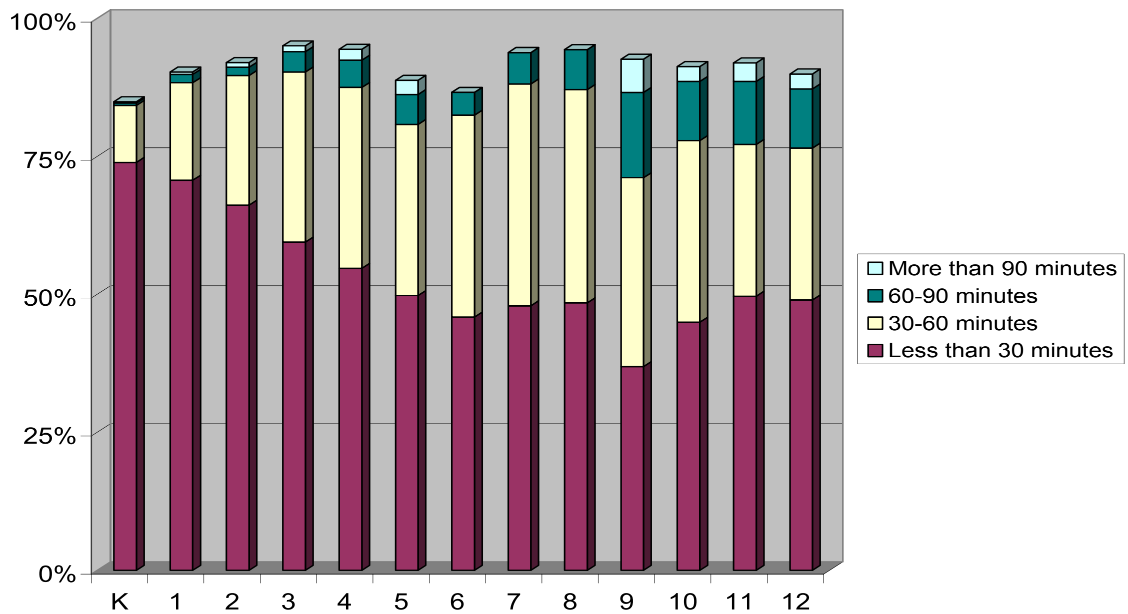


Figure 4.3 shows the pattern of computer usage by grade. The heights of the bars are the total percentage of students using computers, which are consistent with Figure 4.1, allowing for measurement error. Usage increased over most of the elementary grades, fluctuated in middle school and was more consistent in high school. Because most students are using computers by grade three, these data would support beginning electronic testing at that level.

The sharpest difference across grades is the percentage spending less than 30 minutes per day compared to the percentage spending 30 to 60 minutes. While the total percentage of students using computers did not increase after grade three, the amount of time spent by each

[2] The percents for middle school grades may be underestimated due to differences in configurations. See Figure 4.1.

student continued to increase through elementary and middle school. Consequently, while it may be reasonable to begin some form of electronic testing at grade three, the amount of time spent using a computer may not prepare the student for the full NCLB assessment might not be sufficient.

EXPERT PANEL

The questions about where to begin eTesting were also posed to the representatives of nine states who participated in the SC Feasibility Study Expert Panel on March 28, 2007, in Columbia. The specific grades and content areas where eTesting was first implemented varied. In many cases, these were End-of-Course exams. Often they involved smaller initial populations and initially relied on voluntary participation. This allowed the systems and procedures to be tried without immediately needing to confront all the limiting cases of infrastructure and technical support or interfering with high stakes accountability programs.

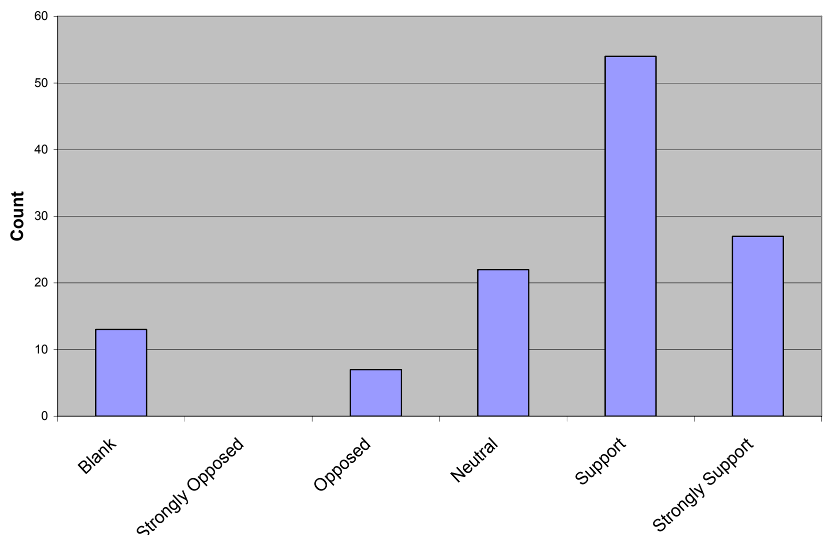
In one state, Computer Skills was chosen because it was a natural fit with the eTesting medium. Another state began with special needs assessments because of the smaller population and the opportunities provided by a variety of built-in aids and accommodations. One state began with grade seven geography because it was a completely new assessment. This avoided the need for any sort of link to an existing assessment or scale or for comparability with the prior year's data. In another case, a direct writing assessment with artificial intelligence (AI) scoring was the initial offering because students were receiving writing instruction on computers and because the scoring approach offered prompt feedback.

The panel members emphasized the importance of beginning slowly and developing the program carefully. It is important to satisfy a need and to obtain the support of the teachers, administrators, and technology staff. Students almost universally reacted positively, often enthusiastically, to eTesting.

Several states tried beginning with voluntary participation to facilitate the start-up, but met with mixed success. The reluctance to participate on a voluntary basis seemed to be a fear of the unknown by administrators and an unwillingness to be on the cutting edge in a high stakes environment. The sole incentive used to get a high level of voluntary participation in Kansas was the promise, and delivery, of diagnostic reports quickly.

To succeed, the eTesting program must build on small, early successes rather than attempt to recover from an initial disaster. The SC program implementation decisions will ultimately be driven by the local situation, but should take into account these experiences of early participants in eTesting.

Figure 4.4: Strength of Support for eTesting



Open-Ended Questionnaire Comments

The informal questionnaires completed by the educators provided several opportunities for respondents to provide comments. The comments received offer intriguing insights for understanding the respondents' thinking about eTesting. Overall, there was strong support for moving to a computerized assessment, with two-thirds of the respondents indicating *support* or *strongly support* (Figure 4.4).

An open-ended question about *barriers to eTesting* was intended to elicit factors that might make eTesting difficult to implement; in contrast, the question about *disadvantages* was seeking reasons SDE might not want to implement it.

The **barriers** mentioned were:

- too few computers,
- inequities across districts,
- too expensive,
- lack of technical support,
- need for professional development for teachers and staff,
- scheduling of testing sessions,
- need for longer testing window, and
- slow or unreliable Internet connections.

Under **disadvantages**, they mentioned the same concerns but also added that eTesting may be:

- poor match with instruction,
- security risk, easier to cheat, and
- unfair to some students.

The security concerns dealt primarily with the arrangement of the computer terminals and preventing one student from seeing another's screen. Computer instruction and labs are often organized to encourage collaboration; thus, there would need to be some adaptation of the physical space. There was little concern expressed about on-line security.

It was suggested that eTesting might be unfair when students do not receive instruction in the same medium. Writing is an obvious example because it requires some practice and is not the same process as with paper and pencil. It is also an issue with math and science, when graphing or manipulatives (e.g., ruler, compass) are involved. With reading, long passages that require scrolling are distractions. In general, recent eTesting platforms incorporate many of the same aids, such as highlighting and crossing out wrong answers. Comparability studies confirm that training and practice on the system are critical.

The second *Fairness* issue was concerned with inequities of resources among schools and among students. However, the survey results suggest students are using computers at all levels. Comparability studies (e.g., Bridgeman et al., 1999) found little effect of differences in computer experience, provided adequate training on the assessment system was provided.

The **advantages** listed are informative. They suggest the expectations that the front line educators have for eTesting, which will need to be met if eTesting is to be successful. Items mentioned by numerous respondents were:

- less costly in the long run,
- less paper without printed materials,
- easier logistics without the receiving and shipping of materials,
- possibly adaptive, and
- very importantly, students like it.

But by far, the number one advantage, listed by almost everyone, was

- student-level diagnostic reports will be in the hands of teachers within days of testing.

Summary

- The educators responding to the questionnaire prefer starting eTesting with reading and math because of the central role these content areas have in the curriculum.
- There was a concern from educators that it may be unfair to assess writing in elementary grades because of the lack of practice composing on computers.
- The Web survey indicated widespread practice with computers including writing, which began early and continued through grade 12.
- The Expert Panel strongly advised to begin small, using volunteer participants to build capacity, confidence, and support.
- The most significant perceived advantage of eTesting was that reports would be received quickly and have the diagnostics necessary to effect instructional improvements.
- The Web survey results of computer exposure and training suggest that any content area could be assessed with eTesting and any grade from grade three on could be included.
- It is critical that the early forays succeed.

SUGGESTIONS FOR IMPLEMENTATION

There are many ways one might begin to implement eTesting and any approach will involve trade-offs. Many of the considerations relate to policy rather than technology or psychometrics. One can argue for beginning with a non-NCLB content area. On the other hand, teachers in their responses to the questionnaire favored the two central NCLB areas of reading and math.

COMPONENT 4 – REFERENCES AND LINKS

American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME) (1999). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

American Psychological Association (1986). *Guidelines for Computer-Based Tests and Interpretations*, Washington, DC: American Psychological Association.

Association of Test Publishers (2002). *Computer Based Testing Guidelines*, York, PA: Association of Test Publishers.

Bennett, R.E. (2003). *Online assessment and the comparability of score meaning.* Princeton, NJ: Educational Testing Service.

Bennett, R.E. (2002a). Inexorable and inevitable: The continuing story of technology and assessment. *Journal of Technology, Learning, and Assessment, 1*(1). Available from http://www.jtla.org.

Bolt, D., & Crawford, R. (2000). *Digital divide: Computers and our children's future*. New York: TV Books.

Bridgeman, B., Bejar, I. I., & Friedman, D. (1999). Fairness issues in a computer-based architectural licensure examination. *Computers in Human Behavior*, 15, 419–440.

Bridgeman, B., Lennon, M.L., & Jackenthal, A. (2001). *Effects of screen size, screen resolution, and display rate on computer-based test performance* (ETS RR-01-23). Princeton, NJ: Educational Testing Service.

Brown-Chidsey, R., & Boscardin, M. L. (1999). *Computers as accessibility tools for students with and without learning disabilities.* Amherst, MA: University of Massachusetts.

Brown-Chidsey, R., Boscardin, M. L., & Sireci, S. G. (1999). *Computers attitudes and opinions of students with and without disabilities.* Amherst, MA: University of Massachusetts.

Brown, P. J., & Augustine, A. (2001). *Screen reading software as an assessment accommodation: Implications for instruction and student performance*. Paper presented at the American Education Research Association Annual Meeting, Seattle, WA, April, 2001.

Cerillo, T.L. & Davis, J.A. (2004). *Comparison of paper-based and computer based administrations of high-stakes, high-school graduation tests.* Paper presented at the Annual Meeting of the American Education Research Association, San Diego, CA.

Choi, S.W. & Tinkler, T. (2002). *Evaluating comparability of paper and computer-based assessment in a K–12 setting.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Donovan, M. A., Drasgow, F., Probst, T. M. (2000). Does computerizing paper-and-pencil job attitude scale make a difference? New IRT analyses offer insight. *Journal of applied Psychology*, 85(2), 305–313.

EdWeek (2003a). *Legal Twists, Digital Turns, 2003 Editorial Projects in Education,* 22(35), pp. 11-16, 16. Retrieved on 02/10/06 from http://counts.edweek.org/sreports/tc03/article.cfm?slug=35impact.h22.

EdWeek (2003b). *Pencils Down: Technology's Answer to Testing*. Retrieved on 02/10/06 from http://counts.edweek.org/sreports/tc03/article.cfm?slug=35exec.h22.

EdWeek (2003). *Tech's Answer to Testing, 2003 Editorial Projects in Education,* 22(35), pp. 8, 16. Retrieved 02/10/06 from http://counts.edweek.org/sreports/tc03/article.cfm?slug=35.h22.

EdWeek (2005). *Testing Computerized Exams. Education Week on the Web.* Retrieved on 02/20/2007 from: http://www.edweek.org/chat/transcript_12_14_2005.html?levelld=2200&levelld=1000.

Ewing, M., Wiley, A. & Gillie, J.M. (2003). *Moving from paper-and-pencil administration to computer-based testing: An investigation of construct equivalence and subgroup differences*. Paper presented at the 2003 NCME Annual Meeting, April 2003, Chicago, IL

Gallagher, A., Bridgeman, B., and Cahalan, C. (2000). *The Effect of Computer-Based Tests on Racial/Ethnic, Gender, and Language Groups (ETS Research Rep. No. 00-8)*. Reviewed 3/10/07 from http://www.ets.org/Media/Research/pdf/RR-00-08-Gallagher.pdf.

Glassnapp, D.R., Poggio, J., Poggio, A., & Yang, X. (2005). *Student attitudes and perceptions regarding computerized testing and the relationship to performance in large scale assessment programs*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal.

Goldberg, A., Russell, M., & Cook, A. (2003). The effect of computers on student writing: A metaanalysis of studies from 1992 to 2002. *Journal of Technology, Learning, and Assessment, 2*(1). Available from http://www.jtla.org.

Higgins, J., Russell, M., & Hoffmann, T. (2005). Examining the effect of computer-based passage presentation on reading test performance. *Journal of Technology, Learning, and Assessment, 3*(4). Retrieved on 3/30/06 from http://www.jtla.org.

Horkay, N., Bennett, R E., Allen, N., Kaplan, B., & Yan, F. (2006). 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(2), Retrieved 1/20/07 from http://www.jtla.org.

Ito, K. & Sykes, R.C. (2004). *Comparability of scores from norm-referenced paper-and-pencil and web-based linear tests for grades 4–12*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Johnson, M. & Green, S. (2004). *On-line assessment: the impact of mode on students' strategies, perceptions, and behaviors.* Paper presented at the annual meeting of the British Educational Research Association, Manchester, Great Britain.

Kim, D. & Huynh, H. (2006). *A comparison of student performance between paper-and-pencil and computer-based testing in four content areas*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Kingston, N.M. (2004). *The comparability of scores of a direct writing measure administered on computer versus administered on paper*. Unpublished report presented to the Maine Department of Education.

Kingston, N.M. (2002). *Comparability of scores from computer- and paper-based administrations for students in grades K-8*. 32nd annual Large-Scale Assessment Conference of the Council of Chief State School Officers, Palm Desert, CA.

Kolen, M. J. (1999-2000). Threats to score comparability with applications to performance assessments and computerized adaptive tests. *Educational Assessment*, 6, 73–96.

Makiney, J. D., Rosen, C., Davis, B. W., Tinios, K. & Young, P. (2003). *Examining the measurement equivalence of paper and computerized job analyses scales.* Paper presented at the 18th Annual Conference of the Society for Industrial and Organizational Psychology, Orlando, FL.

Massachusetts Department of Education (2006). *Report and Analysis of the MCAS Online Pilot Feasibility Study Using the Measured Progress iTest System*.

McCann, R. (2006). The equivalence of online and traditional testing for different subpopulations and item types. *British Journal of Educational Technology*, 37(1), 79–81.

Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin, 114*(3), 449–458.

Merten, T. (1996). A comparison of computerized and conventional administration of the German versions of the Eysenck Personality Questionnaire and the Carroll Rating Scale for Depression. *Personality and Individual Differences*, 20, 281–291.

Murphy, P.K., Long, J., Holleran, T., & Esterly, E. (2000). *Persuasion online or on paper: A new take on an old issue*. Paper presented at the annual meeting of the American Psychological Association, Washington, D.C.

New England Compact (February, 2005) *Computers Improve Students' Writing Performance on Tests.* Education Development Center: Newton, MA. Retrieved May 10, 2007 from http://www.necompact.org/NECompact%20Summaries_0205.pdf

Nichols, L.M. (1996). Pencil and paper versus word processing: a comparative study of creative writing in elementary school. *Journal of Research on Computing in Education, 29,* 59–66.

Nichols, P. & Kirkpatrick, R. (2005). *Comparability of the computer-administered tests with existing paper-and-pencil tests in reading and mathematics tests*. Paper presented at the AERA Division D Graduate Student Seminar, Montreal, Canada.

O'Dwyer, L., Russell, M., Bebell, D., & Tucker-Seeley, K. R. (2005). Examining the relationship between home and school computer use and students' English Language Arts Test Scores. *Journal of Technology, Learning, and Assessment*, 3(3), 4–45.

O'Malley, K. J., Kirkpatrick, R., Sherwood, W., Burdick, H. J., Hsieh, M.C., Sanford, E.E. (2005). *Comparability of a Paper Based and Computer Based Reading Test in Early Elementary Grades*. Paper presented at the AERA Division D Graduate Student Seminar, Montreal, Canada.

Ommerborn, R., & Schuemer, R. (2001). *Using computers in distance study: Results of a survey amongst disabled distance students.* FernUniversitat-Gesamthochschule in Hagen. Retrieved 2/1/2007, from: http://www.fernuni-hagen.de/ZIFF

Paek, P. (2005). *Recent trends in comparability studies* (PEM Research Report 05-05). Available from http://www.pearsonedmeasurement.com/downloads/research/RR_05_05.pdf.

Pearson Educational Measurement. (2002). *Final report on the comparability of computer-delivered and paper tests for Algebra I, Earth Science and English.* Austin, TX: Author.

Pearson Educational Measurement. (2003). *Virginia standards of learning web-based assessments comparability study report – Spring 2002 administration: Online & paper tests*. Austin, TX: Author.

Pearson Educational Measurement (2006). *Comparability analysis of online and paper-and-pencil MEAP*. Iowa City, IA: Author.

Pinsoneault, T. B. (1996). Equivalency of computer-assisted paper-and-pencil administered version of the Minnesota Multiphasic personality Inventory-2. *Computers in Human Behavior*, 12, 291–300.

Poggio, J., Glasnapp, D.R., Yang, X. & Poggio, A.J. (2005b). A comparative evaluation of score results from computerized and paper & pencil mathematics testing in a large scale state assessment program. *The Journal of Technology, Learning, and Assessment, 3*(6), 1–30.

Pommerich, M. (2004). Developing computerized versions of paper tests: Mode effects for passage-based tests. *The Journal of Technology, Learning, and Assessment, 2*(6), 1-44.

Pomplun, M., Frey, S. & Becker, D.F. (2002). The score equivalence of paper and computerized versions of a speeded test of reading comprehension. *Educational and Psychological Measurement, 62*(2), 337–354.

Power, D. E., & Potenza, M. T. (1996). *Comparability of testing using laptop and desktop computers* (RR-96-15). Princeton, NJ: Educational Testing Service.

Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: A comparison of confirmatory factor analysis and item response theory. *Journal of Applied Psychology*, 87 (3), 517–529.

Russell, M. (1999). Testing on computers: A follow-up study comparing performance on computer and on paper. *Education Policy Analysis Archives, 7*(20). Retrieved 2/20/2007, from http://epaa.asu.edu/epaa/v7n20

Russell, M. (2002). It's Time to Upgrade: Tests and administration procedures for the new millennium. *Essays in Education*, 1(1). Available online: http://www.columbiacollegesc.edu/essays/vol.%201%20spring%202002/time_to_upgrade_revised.pdf.

Russell, M. & Hoffmann, T. (2005). Examining the Effect of Computer-Based Passage Presentation on Reading Test Performance. *Journal of Technology, Learning, and Assessment*, 3(4). Retrieved 3/25/07, from http://www.jtla.org

Russell, M. & Plati, T. (2001). *Effects of computer versus paper administration of a state-mandated writing assessment.* Teachers College. Retrieved February, 2005, from http://www.tcrecord.org

Russell M. & Haney. W. (2000a). Bridging the gap between testing and technology in schools. *Education Policy Analysis Archives, 8*(19). Retrieved February 1, 2005, from http://epaa.asu.edu/epaa/v8n19.html

Russell, M. & Haney, W. (1997). Testing writing on computers: An experiment comparing student performance on tests conducted via computer and via paper . *Educational Policy Analysis Archives, 5*(3). Retrieved February 1, 2005, from http://epaa.asu.edu/epaa/v8n19.html

Sandene, B., Horkay, N., Bennett, R., Allen, N., Kaplan, B., and Oranje, A. (2005). Online assessment in mathematics and writing: *Reports from NAEP Technology-Based Assessment Project, Research and Development Series* (NCES 2005-457). U.S. Department of Education, National Center for Educational Statistics. Washington, D.C.: U.S. Government Printing Office. Retrieved on 3/25/07 from http://nces.ed.gov/nationsreportcard/pdf/studies/2005457_1.pdf.

Schaeffer, G. A., Bridgeman, B., Golub-Smith, M. L., Lewis, C., Potenza, M. T., & Steffen, M. (1998). *Comparability of paper-and-pencil and computer adaptive test scores on the GRE General Test (RR-98-38)*. Princeton, NJ: Educational Testing Service.

Sim, G. & Horton, M. (2005) *Performance and Attitude of Children in Computer Based Versus Paper Based Testing.* World Conference on Educational Multimedia, Hypermedia and Telecommunications, Montreal.

Stocking, M. (1997) Revising item responses in computerized adaptive tests: a comparison of three models. *Applied Psychological Measurement. 21*, 129–142.

Taylor, C., Jamieson, J., Eignor, D., & Kirsch, I. (1998). *The relationship between computer familiarity and performance on computer-based TOEFL test tasks (TOEFL-RR-61).* Princeton, NJ: Educational Testing Service.

Trotter, A. (2001). Testing computerized exams. *Education Week*, 20 (37) 30-35. Retrieved March, 2002, from: http://www.edweek.org/ew/westory.cfm?slug=37online.h20

U.S. Department of Commerce. (2002). *A nation online: How americans are expanding their use of the internet*. Washington, DC: Author. Retrieved April 19, 2002, from http://www.ntia.doc.gov/ntiahome/dn/nationonline_020502.htm

U.S. Department of Education, Office of Planning, Evaluation and Policy Department U.S. Department of Education (2007). *State Strategies and Practices for Educational Technology: Volume I—Examining the Enhancing Education Through Technology Program*. Washington, DC: Author.

Vispoel, W. P., Clough, S. J., Bleiler, T., Hendrickson, A. B., & Ihrig, D. (2002) Can examinees use judgments of item difficulty to improve proficiency estimates on computerized adaptive vocabulary tests? *Journal of Educational Measurement, 39,* 311–330.

Vispoel, W., Hendrickson, A. B., & Bleiler, T. (2000). Limiting answer review and change on computerized adaptive vocabulary test: Psychometric and attitudinal results, *Journal of Educational Measurement,* 37, 21–38.

Wang, S. (2004). *Online or paper: Does delivery affect results? Administration mode comparability study for Stanford Diagnostic Reading and Mathematics Tests.* San Antonio, Texas: Harcourt.

Wang, S., Young, M., and Brooks, T. (2004). *Administration mode comparability study for Stanford Diagnostic Reading and Mathematics Tests*. San Antonio, Texas: Harcourt.

Way, W. D., Davis, L. L., & Fizpatrick, S. (2006). *Practical Questions in Introducing Computerized Adaptive Testing for K-12 Assessments*. Retrieved 3/26/07 from http://www.pearsonedmeasurement.com/research/research.htm#

Williams, S. (n.d.) *Online Testing*. Retrieved 3/26/2007 from www.indiana-etc.org/pdfs%5COnline-Testing.pdf

Wissick, C. A., & Gardner, J. E. (2000). Muttimedia or not to multimedia? That is the question for students with learning disabilities. *Teaching Exceptional Children*, 32 (4), 34–43.

Yang, X., Poggio, A. Glasnapp, D., & Poggio, J. (2007) *Score comparability between computerized and paper-and-pencil linear tests.* Unpublished paper.

Zandviliet, D., & Farragher, P. (1997). A comparison of computer-administered and written tests. *Journal of Research on Computing in Education*, 29(4), 423–438.

COMPONENT 5 – FEASIBILITY OF INCLUSION OF CONSTRUCTED-RESPONSE ITEMS AS PART OF THE WRITING ASSESSMENT AND FEASIBILITY OF MOVEMENT OF THE WRITING ASSESSMENT TO A SEPARATE ADMINISTRATION

Multiple-choice (MC) items, until very recently, were often termed *objective* items. This language arose because, once the item was developed and administered, anyone, even a computer, could score it with perfect consistency and without ambiguity. In contrast, open-ended, short answer, and essay items seemed to lack this objectivity. Thus, the lingering doubts about the objectivity and appropriateness of these items, collectively called *Constructed Response* (CR) in this discussion. The direct assessment of writing is an important special case of the application of non-multiple-choice, non-dichotomous items.

Constructed Response items provide integer scores ranging from zero to some value greater than one, often four or five. Summary scoring generally gives the same value to each point gained from a CR item as is given to each point from a correct response to a multiple choice item. Consequently, a five-point CR item is assumed equal in weight to five one-point MC items.

Educators find CR attractive for a number of reasons, but the primary appeal seems to be the perception that higher order skills can only be tapped into with CR tasks. These tasks appear more relevant and less contrived than MC often do. The same task can cover a wider range of student proficiency by giving the high performing students the opportunity to excel and the students at the lower end the opportunity to demonstrate some proficiency. The significance of CR is reflected by their increasing role in the NAEP assessment and the college admission exams.

CR are most valuable to teachers when they are given the actual student responses. It is then that they can begin to understand what the student is thinking and what the appropriate next steps might be in regard to targeting instruction.

However, CR are expensive; first, because of the effort required for developing the item and the rubric and for hand scoring the responses. They are also expensive in terms of the time required for the student to respond. Typically, ten to fifteen or more minutes are allowed for responding to a five-point CR item. In the same length of time, a student could respond to ten or twelve MC. In simple terms of a cost-benefit comparison, the students' time could be used more efficiently if 12 one-point MC items were presented rather than one 5-point CR.

Final reporting is often delayed by several weeks to allow time to complete the hand scoring process. While it is difficult to assign a dollar value, student test results that are available shortly after testing would seem to have much greater value than results that may not be available until the beginning of the next school year.

Security issues are somewhat different for CR than for MC. In general, CR are more memorable, perhaps because they occupy so much of the student's time. This would imply CR have a shorter useful life than MC because they become known to students and teachers much faster. It has been suggested that exposure is less critical with CR because, even if the students have practiced the specific task, they still need to perform the task on the test, and all the practice just helped them achieve proficiency. It is difficult to defend this lack of concern in a high stakes environment where some teachers and students may be sufficiently motivated to memorize the response in its entirety in advance.

This section will address two technical issues associated with the use of CR items and will leave the policy issues just mentioned to another forum. Question I is, what do CR tasks add to the precision of the measurement regardless of the mode of administration? This discussion applies equally well to paper-and-pencil (P&P) as to electronic testing (eTesting).

Question II is, what are the options to expedite scoring of CR in the context of electronic testing? This discussion will explore various options using human raters as well as computerized scoring. The human options include split testing windows and distribution of the CR to online raters via the World Wide Web. Currently marketed computerized approaches to automated scoring include regression, Bayesian, artificial intelligence, and latent semantic analysis models.

Question I: Reliability, Precision, and Consistency

Reliability, precision, and consistency are three views of the same issue. *Reliability* is an index of how well the test functions for this population. *Precision* is a measure of how accurately an individual student is measured. *Consistency* is the degree to which alternate approaches lead to the same decisions.

Reliability

Reliability is an established concept from true score theory that is useful as an overall summary of the *quality* of a test. It is sometimes described as the correlation between the scores on a test and the scores on a hypothetical retest with a parallel form. The higher the correlation, the more *reliable* the scores are.

Reliability is generally computed as the ratio of the variance associated with the measure to the total variance of the observation. The difference between the two is the variance associated with measurement error.

1. $r = \dfrac{\sigma_M^2}{\sigma_M^2 + \sigma_e^2} = \dfrac{\sigma^2 - \sigma_e^2}{\sigma^2}$, where σ_M^2 is the (generally unknown) variance of the measures (or true scores), σ_e^2 is the standard error of measurement squared, and $\sigma^2 = \sigma_M^2 + \sigma_e^2$ is the total variance of the observed scores.

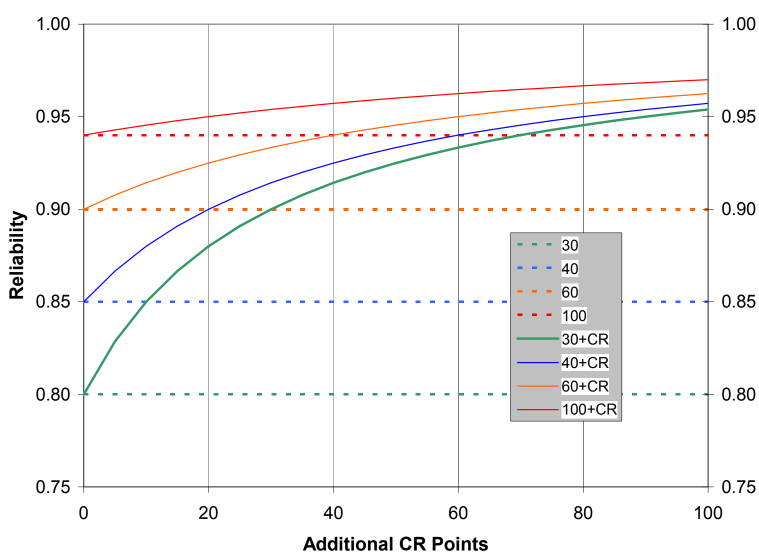
The reliability coefficient will be nearest its maximum value of one when the variance of the measure is large compared to the error of the measurement. In practical terms, this means that the students in the population cover a wide range of proficiency and that the test is very precise. A *precise* test requires many well-crafted items.

Assuming consistently high quality items, reliability is primarily a function of test length, i.e., the number of possible points. A useful rule of thumb for an initial estimate of reliability for use in test design is:

2. $r = \dfrac{L-6}{L}$, where L is the total number of points.

This simple expression implies that a thirty-point test could be expected to have a reliability of about 0.80; a forty-point test, about 0.85; and a sixty-point test, about 0.90.

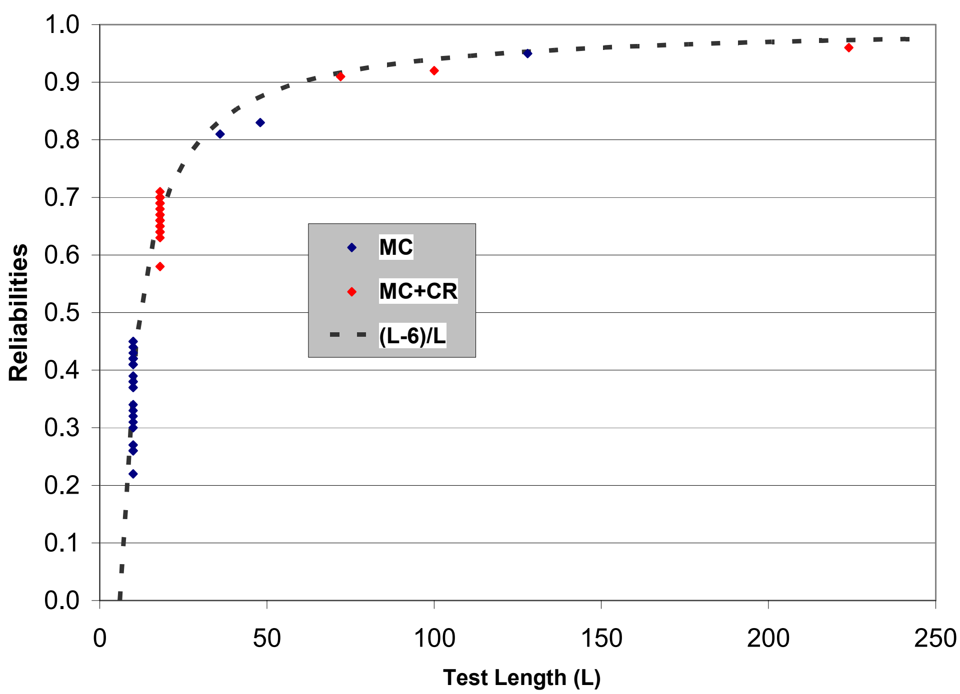
Figure 5.1: Hypothetical Reliability Gains from Adding Points



Under these parameters, Figure 5.1 illustrates the relationship between test length and reliability. The horizontal dotted lines represent the typical reliabilities for tests with 30, 40, 60 and 100 points, respectively. The curved solid lines indicate the increased reliability due to adding points to each of the base tests. This display implies that a test with 30 MC items and 70 CR points is equivalent to a 100-item MC test. This is the relationship that one would expect if all points, CR or MC, are equivalent, and it will be used as a benchmark to evaluate later results.

Figure 5.2 summarizes the results of the analysis of 34 separate tests using both MC and CR. Tests of reading, mathematics, and writing were included, and were drawn from assessments in three different states. The writing assessment included both MC and a standard writing prompt. The dotted line in Figure 5.2 represents the (L-6)/L rule of thumb.

Figure 5.2: Observed (Winsteps) Reliabilities for 34 Tests



Tables 5.1 and 5.2 present similar reliabilities for South Carolina in-state reading and math assessments[1] for grades 3 through 8. Reliabilities are given for the intact assessment, for the MC portion of the assessment alone, and for versions including CR or CR plus ER (*extended responses* to the writing prompt) with random deletions of MC to maintain the total number of points. The CR had little or no effect on the reliability compared to the MC-only version. There is a *suggestion*, particularly in grade 7, that adding the ER task could increase the overall reliability, although it did reach the level of statistical significance.

The conclusion to be drawn is that these data support the conclusion that a point from a CR is equivalent to a point from an MC. Both the MC and the CR closely approximate the benchmark expectations. Consequently, a decision about whether or not to use CR will be decided not on the relative *quality* of the points but rather the total costs associated with the CR and their value for purposes other than simple measurement.

[1] These results are based on a sample of approximately 7,000 students for each grade, so the results will not be identical to the state-wide results. The sample was originally designated as the calibration sample for the operational assessment.

Table 5.1: Reliability Comparison, Grade 3–8 Reading

	# Items				# Score Points	All Items	Item Samples for Reliability*				
	MC	CR	ER	Total			Sample 1	Sample 2	Sample 3	Sample 4	Sample 5
Gr. 3 Reading N=6934											
MC Only	31	0	0	31	31	0.83					
MC + CR	27	2	0	29	31		0.81	0.81	0.81	0.81	0.81
MC + CR + ER	12	2	4	18	31		0.81	0.82	0.81	0.81	0.75
Full Test	31	2	4	37	50	0.88					
Gr. 4 Reading N=6948											
MC Only	34	0	0	34	34	0.82					
MC + CR	29	2	0	31	34		0.81	0.81	0.80	0.81	0.81
MC + CR + ER	14	2	4	20	34		0.81	0.81	0.81	0.81	0.81
Full Test	34	2	4	40	54	0.87					
Gr. 5 Reading N=6918											
MC Only	33	0	0	33	33	0.81					
MC + CR	27	3	0	30	33		0.81	0.81	0.81	0.81	0.81
MC + CR + ER	12	3	4	19	33		0.83	0.83	0.82	0.81	0.82
Full Test	33	3	4	40	54	0.88					
Gr. 6 Reading N=6766											
MC Only	46	0	0	46	46	0.85					
MC + CR	37	3	0	40	46		0.86	0.86	0.86	0.86	0.86
MC + CR + ER	22	3	4	29	46		0.87	0.88	0.88	0.89	0.88
Full Test	46	3	4	53	60	0.92					
Gr. 7 Reading N=6765											
MC Only	44	0	0	44	44	0.81					
MC + CR	33	4	0	37	44		0.81	0.82	0.82	0.82	0.83
MC + CR + ER	18	4	4	26	44		0.86	0.86	0.86	0.86	0.86
Full Test	44	4	4	52	70	0.90					
Gr. 8 Reading N=6846											
MC Only	49	0	0	49	49	0.85					
MC + CR	36	4	0	40	49		0.85	0.85	0.85	0.85	0.84
MC + CR + ER	21	4	4	29	49		0.87	0.87	0.86	0.87	0.87
Full Test	49	4	4	57	77	0.91					

Table 5.2: Reliability Comparison, Grade 3–8 Mathematics

	# Items					Item Samples for Reliability*				
	MC	CR	Total	# Score Points	All Items	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5
Gr. 3 Mathematics N=7035										
MC Only	40	0	40	40	0.85					
MC + CR	34	2	36	40		0.85	0.85	0.85	0.85	0.85
Full Test	40	2	42	46	0.87					
Gr. 4 Mathematics N=7034										
MC Only	41	0	41	41	0.87					
MC + CR	36	2	38	41		0.85	0.85	0.85	0.85	0.85
Full Test	41	2	43	45	0.88					
Gr. 5 Mathematics N=7001										
MC Only	40	0	40	40	0.86					
MC + CR	34	2	36	40		0.87	0.87	0.86	0.86	0.86
Full Test	40	2	42	46	0.88					
Gr. 6 Mathematics N=6875										
MC Only	54	0	54	54	0.88					
MC + CR	48	2	50	54		0.89	0.88	0.88	0.89	0.88
Full Test	54	2	56	60	0.89					
Gr. 7 Mathematics N=6797										
MC Only	54	0	54	54	0.85					
MC + CR	48	2	50	54		0.85	0.85	0.85	0.85	0.85
Full Test	54	2	56	60	0.86					
Gr. 8 Mathematics N=6880										
MC Only	65	0	65	65	0.91					
MC + CR	58	2	60	65		0.91	0.91	0.91	0.91	0.91
Full Test	65	2	67	72	0.92					

PRECISION

The concept of precision addresses the issue of measurement accuracy directly and will be defined here as the *standard error of measurement*. Mathematically, it is closely associated with reliability. As shown in expression 1 on page 5-4, for a given population of students, the assessment with the lower standard error will have the higher reliability.[2]

The standard error is generally considered a more appropriate index for evaluating a standards-based, criterion-referenced assessment. A standards-based assessment is primarily concerned with inferences about the status of individuals with respect to an external criterion. Consistent with this intent, the standard error of measurement describes how accurately the individual has been measured. It can be meaningfully applied to interpret the score of a single student. In contrast, reliability describes how consistently the instrument orders a population of students.

The data in Figure 5.3 were taken from a South Carolina grade three reading test. Six scores were generated for each student after the fact. Each score involved 62 total points. The first used all 62 MC items administered. The others used 54 MC items and four two-point CR items. To create forms with equal numbers of points, 8 MC items were deleted at random for each score.

The metric used in the figure is the Rasch logit metric, which is transformed into a more user-friendly scale score prior to reporting. For purposes of these comparisons, the choice of metric is immaterial.

Other than slight differences in the average item difficulty, which manifests itself in the slight variation in the logit scores at the extremes, there is no meaningful difference in the

[2] This assertion presumes the raw scores, or number correct, have been transformed into a more appropriate measurement scale.

precision of the measurement with and without the CR items. Randomly replacing eight MC with four two point CR on a 62-point test is admittedly a very modest alteration of the test.

Figure 5.3: Standard Error of Measurement with and without CR

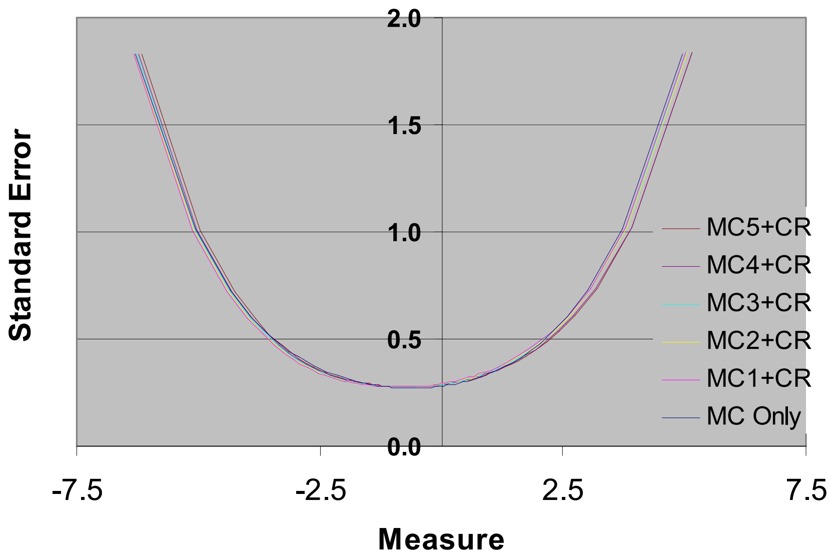
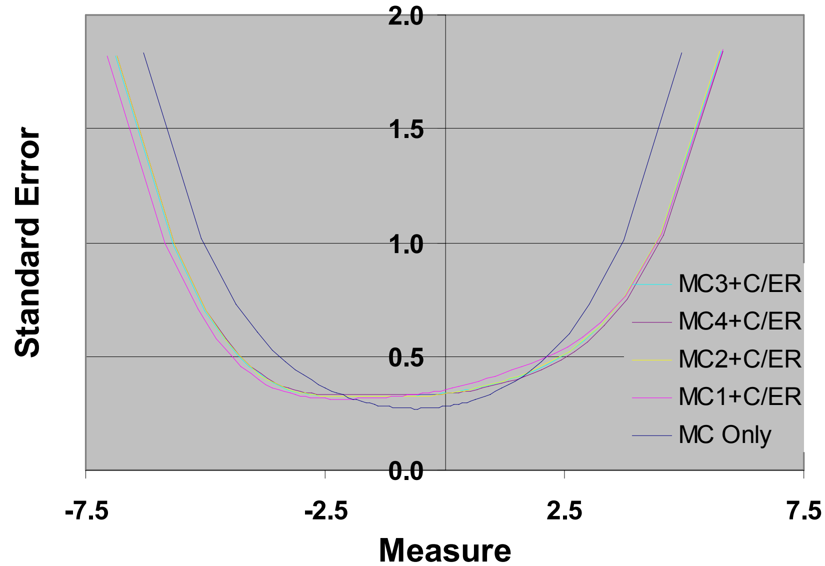


Figure 5.4 presents the results of a more severe alteration to the same assessment. These data include an extended response (ER) exercise that was scored on multiple domains. There was a maximum of 30 points possible for the ER task. In addition to the same MC-only version as before, the four variations shown here have 24 MC, 8 CR and 30 ER points.

Including the ER task widened the effective range of the test at the cost of some precision in the center. While the standard errors of measurement (SEM) are higher in the center, they are lower for the more extreme scores. This effect is produced by the location of the score thresholds for the ER. Scores of one represent very low proficiency while scores of three or four represent very high proficiency. This reflects the task and the scoring rubric and does not indicate higher or lower quality for the ER points compared to the MC. Again, it made no practical difference which MC items were randomly dropped.

As was the case earlier with reliability, this investigation of precision does not provide a compelling argument for or against using CR/ER tasks. There is a design issue that the MC versions were more sharply focused in a narrow range of proficiency while the ER version would be more effective over a broader range. Which design is optimal depends on the purpose for which the scores are intended.

Figure 5.4: Standard Error of Measurement when CR and ER are included



DECISION CONSISTENCY

For a high stakes assessment in which the intent is to classify each student into one of two or more performance categories, the most relevant criterion for judging the contribution of CR may be *decision consistency*. The fundamental question is whether the assessment including CR classifies students the same way as the assessment without CR. While there have been a number of indices proposed to quantify decision consistency, the most natural and understandable statistics are counts of students classified into the same categories.

Table 5.3: Consistency Indices for Performance Levels

MC-only Tests vs. Full Length Tests					
		Two Performance Levels		Four Performance Levels	
Grade	Subject	Proportion of Agreement	Kappa	Proportion of Agreement	Kappa
3	ELA	0.94	0.88	0.84	0.76
	Mathematics	0.95	0.89	0.89	0.84
4	ELA	0.92	0.84	0.84	0.77
	Mathematics	0.97	0.94	0.92	0.90
5	ELA	0.91	0.81	0.84	0.75
	Mathematics	0.97	0.92	0.92	0.88
6	ELA	0.91	0.77	0.79	0.69
	Mathematics	0.97	0.93	0.92	0.88
7	ELA	0.90	0.71	0.80	0.68
	Mathematics	0.97	0.93	0.92	0.88
8	ELA	0.90	0.73	0.80	0.69
	Mathematics	0.98	0.95	0.94	0.90

Table 5.3 displays some summary statistics for recent South Carolina assessments. In preparing this analysis, two scale scores were computed: one using all available MC items, and a second using all available CR with the MC. Then the scale scores were used to classify each student into a two-way table. Ideally, both scale scores should place each student in the same performance level. These results are based on a sample of approximately 7,000 students for each grade that was selected as the calibration sample so the results will not be identical to the state-wide results.

In general, the decision consistency between the two scales is good. For the simpler case of two levels (proficient or not), there was over 90% agreement in all cases. The results for mathematics are more consistent than for ELA and the results for the lower grades are more consistent than for the higher grades. Tables 5.4 and 5.5 contain detailed counts and percentage for all grades and content areas. The level of agreement from the two measures reflects the variation in scale scores associated with one or two raw score points.

Table 5.4: Student Performance Level Placement — Full Test versus MC Only

	Grade		Language Arts MC Only Version					Math MC Only Version				
			B.B.	Basic	Prof.	Adv.	Total	B.B.	Basic	Prof.	Adv.	Total
Full Test	Three	Below Basic	**885**	160	1	0	1046	**1289**	82	0	0	1371
		Basic	136	**1921**	192	0	2249	91	**2962**	142	0	3195
		Proficient	0	238	**2637**	121	2996	0	199	**1184**	260	1643
		Advanced	0	0	256	**388**	644	0	0	2	**824**	826
		Total	1021	2319	3086	509	**6935**	1380	3243	1328	1084	**7035**
	Four	Below Basic	**1366**	134	0	0	1500	**1532**	21	0	0	1553
		Basic	213	**2063**	383	0	2659	69	**2305**	177	0	2551
		Proficient	0	153	**2187**	86	2426	0	38	**1493**	98	1629
		Advanced	0	0	128	**235**	363	0	0	134	**1167**	1301
		Total	1579	2350	2698	321	**6948**	1601	2364	1804	1265	**7034**
	Five	Below Basic	**1390**	179	0	0	1569	**1592**	112	0	0	1704
		Basic	147	**2506**	366	0	3019	76	**2695**	144	0	2915
		Proficient	0	240	**1784**	69	2093	0	102	**1099**	72	1273
		Advanced	0	0	133	**104**	237	0	0	79	**1030**	1109
		Total	1537	2925	2283	173	**6918**	1668	2909	1322	1102	**7001**
	Six	Below Basic	**2084**	236	1	0	2321	**1391**	169	0	0	1560
		Basic	271	**2015**	258	0	2544	76	**2783**	102	0	2961
		Proficient	0	358	**1005**	83	1446	0	107	**1267**	78	1452
		Advanced	0	0	235	**220**	455	0	0	40	**862**	902
		Total	2355	2609	1499	303	**6766**	1467	3059	1409	940	**6875**
	Seven	Below Basic	**1756**	331	1	0	2088	**1608**	208	0	0	1816
		Basic	230	**2630**	187	0	3047	49	**2917**	116	0	3082
		Proficient	0	478	**944**	18	1440	0	69	**1023**	21	1113
		Advanced	0	0	142	**48**	190	0	0	93	**693**	786
		Total	1986	3439	1274	66	**6765**	1657	3194	1232	714	**6797**
	Eight	Below Basic	**1695**	222	2	0	1919	**2235**	223	0	0	2458
		Basic	277	**2609**	271	1	3158	13	**2904**	33	0	2950
		Proficient	0	409	**1025**	70	1504	0	84	**717**	58	859
		Advanced	0	0	128	**137**	265	0	0	25	**588**	613
		Total	1972	3240	1426	208	**6846**	2248	3211	775	646	**6880**

Table 5.5: Percentages of Student Performance Level Placement — Full Test versus MC Only

			Language Arts MC Only Version					Math MC Only Version				
	Grade		B.B.	Basic	Prof.	Adv.	Total	B.B.	Basic	Prof.	Adv.	Total
Full Test	Three	Below Basic	**12.8%**	2.3%	0.0%		15.1%	**18.3%**	1.2%			19.5%
		Basic	2.0%	**27.7%**	2.8%		32.4%	1.3%	**42.1%**	2.0%		45.4%
		Proficient		3.4%	**38.0%**	1.7%	43.2%		2.8%	**16.8%**	3.7%	23.4%
		Advanced			3.7%	**5.6%**	9.3%			0.0%	**11.7%**	11.7%
		Total	14.7%	33.4%	44.5%	7.3%	**100.0%**	19.6%	46.1%	18.9%	15.4%	**100.0%**
	Four	Below Basic	**19.7%**	1.9%			21.6%	**21.8%**	0.3%			22.1%
		Basic	3.1%	**29.7%**	5.5%		38.3%	1.0%	**32.8%**	2.5%		36.3%
		Proficient		2.2%	**31.5%**	1.2%	34.9%		0.5%	**21.2%**	1.4%	23.2%
		Advanced			1.8%	**3.4%**	5.2%			1.9%	**16.6%**	18.5%
		Total	22.7%	33.8%	38.8%	4.6%	**100.0%**	22.8%	33.6%	25.6%	18.0%	**100.0%**
	Five	Below Basic	**20.1%**	2.6%			22.7%	**22.7%**	1.6%			24.3%
		Basic	2.1%	**36.2%**	5.3%		43.6%	1.1%	**38.5%**	2.1%		41.6%
		Proficient		3.5%	**25.8%**	1.0%	30.3%		1.5%	**15.7%**	1.0%	18.2%
		Advanced			1.9%	**1.5%**	3.4%			1.1%	**14.7%**	15.8%
		Total	22.2%	42.3%	33.0%	2.5%	**100.0%**	23.8%	41.6%	18.9%	15.7%	**100.0%**
	Six	Below Basic	**30.8%**	3.5%	0.0%		34.3%	**20.2%**	2.5%			22.7%
		Basic	4.0%	**29.8%**	3.8%		37.6%	1.1%	**40.5%**	1.5%		43.1%
		Proficient		5.3%	**14.9%**	1.2%	21.4%		1.6%	**18.4%**	1.1%	21.1%
		Advanced			3.5%	**3.3%**	6.7%			0.6%	**12.5%**	13.1%
		Total	34.8%	38.6%	22.2%	4.5%	**100.0%**	21.3%	44.5%	20.5%	13.7%	**100.0%**
	Seven	Below Basic	**26.0%**	4.9%	0.0%		30.9%	**23.7%**	3.1%			26.7%
		Basic	3.4%	**38.9%**	2.8%		45.0%	0.7%	**42.9%**	1.7%		45.3%
		Proficient		7.1%	**14.0%**	0.3%	21.3%		1.0%	**15.1%**	0.3%	16.4%
		Advanced			2.1%	**0.7%**	2.8%			1.4%	**10.2%**	11.6%
		Total	29.4%	50.8%	18.8%	1.0%	**100.0%**	24.4%	47.0%	18.1%	10.5%	**100.0%**
	Eight	Below Basic	**24.8%**	3.2%	0.0%		28.0%	**32.5%**	3.2%			35.7%
		Basic	4.0%	**38.1%**	4.0%	0.0%	46.1%	0.2%	**42.2%**	0.5%		42.9%
		Proficient		6.0%	**15.0%**	1.0%	22.0%		1.2%	**10.4%**	0.8%	12.5%
		Advanced			1.9%	**2.0%**	3.9%			0.4%	**8.5%**	8.9%
		Total	28.8%	47.3%	20.8%	3.0%	**100.0%**	32.7%	46.7%	11.3%	9.4%	**100.0%**

Summary

Overall there appears to be no conclusive evidence based on the precision of the measures for or against the inclusion of CR tasks in the assessment.

1. The estimated reliabilities behave consistently whether the additional points are generated by MC or by CR tasks.
2. The standard errors of measurement at the same location on the scale score metric are comparable regardless of the type of item.
3. The consistency of the classification decisions is reasonable given the magnitude of the standard errors of the measures. The level of agreement is very similar to what one would expect if the students were retested with the same item type.

From a strictly measurement perspective and given the cost and effort associated with CR, one would interpret these results to mean that the assessment would be more *efficient* and more economical if CR were not included. More information, in the statistical sense[3], about the student can be obtained in the same amount of testing time at less cost with an all MC test.

However, from an educational perspective, considering the possible impacts on instruction and the value attached to the results by front-line educators and the public, there are persuasive validity arguments for including CR. The results given here are then taken to mean that there is no psychometric reason not to use the CR. Many assessment programs have concluded that CR are well worth the added cost.

[3] *More information, in the statistical sense*, means a smaller standard error of measurement. It does not mean that more is known about the student in general in the popular sense of more information as more data.

Question II: Strategies for Scoring Constructed Response Tasks

Much of the cost associated with constructed response (CR), both in terms of dollars and delays, is related to the scoring process. Historically, the student responses have been individually hand scored against a rubric by carefully trained readers. This process has been found to be reliable and valid over the years for a wide range of content areas, scoring models, and rubrics.

HAND SCORING BY HUMAN READERS

One strategy for dealing with scoring CR in electronic testing (eTesting) is similar to current practice. It would involve transmitting the responses electronically to a panel of trained readers, who then score the responses in much the same manner as they have been, either on-screen or hard-copy. The process would be expedited by the electronic submission, thus avoiding the delays associated with shipping the materials and capturing the data. It does not avoid the delay of the actual scoring. For application in large scale assessment, electronic transmission does not avoid the time and effort associated with sending the results back to the correct school and student.

MIT, in partnership with several other universities, employs this type of strategy as part of its *IMOAT* system for its Freshman Placement Test. *IMOAT* was designed to manage the freshman essay. Its primary strengths are the flexibility that students are allowed and the accessibility by both the student and the advisor to high-quality, detailed feedback. It is also considered a better assessment because the students are allowed several days to research and prepare their essays. While it is neither designed for nor necessarily appropriate for large-scale assessment, it does demonstrate the efficacy of electronic submission and annotated hand scoring either by hard-copy or on-screen.

The Educational Testing Service Online Scoring Network (OSN) is currently being used to score essays for several testing programs, including the SAT[4]. This system distributes essays to trained raters over the World Wide Web. The raters work on somewhat flexible schedules at their own locations using their own computers. All training, quality control and monitoring are conducted online, although scoring supervisors can be available by phone. ETS reports a capacity of 64,000 essays per day.

Further evidence that electronic distribution is viable is provided by numerous states that are currently using it in some form. Indiana, Kentucky, Michigan, Oregon, Texas, and Wyoming are involved with some aspect of it in their electronic testing. See Component 1 for a more detailed discussion of what the states are doing.

SPLIT WRITING TEST WINDOWS

A strategy under consideration to avoid delays in reporting uses an earlier testing window for the writing task. This approach can be used with either paper-and-pencil or electronic tests. It probably is not appropriate for content areas other than writing because it is difficult to logistically separate the ER and MC portions of the test in other areas and to merge the results after testing.

There are additional logistical problems, and related expenses, associated with split windows. It is an extra test administration with all the related organizational details and disruptions for schools and districts. It would entail an additional round of shipping and receiving for the districts and the contractor. Depending on reporting protocols, the contractor might also need to establish a method to collate the results from the split administrations for each student and to provide the resources needed for the collation if the separate windows were not

[4] Currently, under a contract with Pearson Educational Measurement.

self-contained assessments. This is not an issue for writing if there are no MC items included in the writing score and if the writing results are recorded and reported separately from the other content areas.

As a matter of course, there will inevitably be some number of student records from each administration that cannot be matched if joint reporting is involved. In some instances, the lack of a match will be because the student changed schools or districts and entered the identifying information incorrectly and inconsistently. It should be possible, with sufficient time and resources, to resolve these cases completely. In the worst case, it will be a hand search, perhaps involving school personnel.

In some instances, a non-match will result because there is no matching record; the student was absent with no make-up for one of the administrations. This situation requires establishing appropriate policies, which need to cover the various possible causes of the missing data. In principle, this is no different than a student failing to complete a portion of an exam administered in a single window. It could be treated in several ways:

- the student is not considered to have attempted the exam at all (non-attempt) and no score is reported;
- the student receives no credit for the portion not taken, the scale score is computed based on the entire test, and the scale score is generally underestimated; or
- the student is scored only on the portion that was taken and the scale score is *unbiased* but based on a reduced test.

The last option is often recommended for measurement purposes but assumes that the missing data is random and did not occur because the student would have done poorly on the omitted portion. Safeguards would need to be put in place to ensure that schools, or students, do not try to manipulate the process by having their poor writers omit the ER portion of the exam. For an accountability system, this may force the state to adopt the policy of counting the omits as zero scores.

There is also a technical, perhaps philosophical, issue that deserves consideration. In the case of the split windows, there is the assumption that the two occasions are measuring the same variable and that the students have not changed their locations (on the construct) in the interim. While this is problematic because there should be relevant instruction and learning between the windows, it does not adversely affect the measurement if all students change equally. The tasks in the early window will simply appear more difficult than they would have if they had been tested in the late window.

Shifting the timing of any assessment will have implications for equating. To be appropriate for the link, an item's calibrations must come from administrations given at the same time in the school year. When items are shifted by a significant portion of the year, they should be recalibrated for that time period and care taken with any subsequent equating.

Most of these considerations will not be an issue for the direct writing assessment in a separate window if:

- there are no MC items that contribute to the writing score,
- writing is not combined with reading,
- writing results are distributed on separate reports from the other content areas, and
- writing status is not used in combination with any other content area status for graduation or accountability decisions.

Real Time Human Scoring

One major attraction of electronic testing is the capability to report the student's results immediately after the responses are submitted. This, of course, presumes that the items have all been calibrated and performance standards established prior to testing. Because the hand-scoring process is laborious and time-consuming, it can delay reporting of the assessment results for a period of weeks. This delay defeats a principle advantage of eTesting.

In the strictest sense, *real time* implies that scoring takes place while the student is still seated at the computer. Because it only takes a few minutes to score any one response, *real time* scoring is possible if:

1. the student's response is transmitted immediately,
2. a trained rater attends to the submission immediately, and
3. the student can be productively occupied for those few minutes.

First, for the student's response to be transmitted immediately, there must be sufficient *bandwidth* throughout the system to handle all the responses that may be submitted at any moment. The definitive answer to this question depends in part on the Internet capacity of the schools and also how many students might be online and submitting responses simultaneously. Those numbers can be established. However, the final answer will depend on the tests and how they are organized. And, to some extent, it will depend on how much students vary in the time they take to respond to each item. Definitive answers and optimal strategies will require direct, hands-on experience.

Second, even with the current approach to scoring CR, delays in reporting could be eliminated with a large enough pool of raters. There are, of course, economic constraints.

- The cost of recruiting and training a rater is fixed: The cost per student therefore will be lowered by having each rater read more papers.

- Building and equipping a scoring center is expensive: Operating a small facility for a long period of time is more economical than operating a large facility for a short time.
- The number of qualified, trainable raters is limited: The pool of potential raters is often a limitation and typically is a major consideration when siting a scoring center.

Using the World Wide Web to distribute the responses electronically would effectively eliminate the last two constraints. The ETS Online Scoring Network has few regional qualifications[5] for its raters and requires all raters to provide their own sites and their own computers. Security is, of course, a serious concern, but ETS feels its system has addressed the issue appropriately for some high stakes applications in college admission, certification, and licensing.

Finally, occupying the student productively during scoring is manageable but requires some consideration of what is an appropriate use of the student's time or the computer. Organizing the assessment so some MC remain after the CR are submitted is a possible strategy. Another approach might be to suggest the student review the MC responses while waiting for the CR to be scored. Both of these strategies would exclude the opportunity for the student to return to the CR, which may or may not be desirable. Alternatively, there might be administrative details or an opinion survey to attend to.

This entire discussion begs the question of whether it is appropriate to present the assessment results to the student immediately. A much less stringent definition of *real time* might be after lunch or overnight. This would greatly relax the capacity demands on all components: bandwidth, raters, and students. For purposes of large-scale assessment or even formative assessment, this time frame may be acceptable.

[5] Individual clients may specify in-state raters for validity reasons or may exclude in-state for security reasons.

One might also define *real time* as completing the scoring of all CR within a day or two of the close of the testing window. Scoring could begin as soon as any student submitted a CR response. Currently, scoring does not begin until all students have been tested and the materials have been returned and processed at the scoring site. And, often, scoring cannot begin until the *range-finding* process has been completed.

This schedule lacks the dramatic impact of instantaneous feedback but would provide the results in a functional time frame for the teachers and administrators to act upon. It would allow some summary, group-level results to be included. The primary advantage is operational; that is, it would dramatically level the scoring load so that the staffing would need to be able to process the average, not the peak, demand.

Under any of the definitions, real-time scoring requires the CR tasks to have been previously *calibrated*, performance levels established, and the raters trained. Psychometric calibration requires the items to have been tried out on a significant sample of students. Typically, this means a minimum of 1,000 students and often substantially more depending on the psychometric model being used or if they are to be checked for differential functioning in subpopulations of students. These preliminary data need to be obtained under similar conditions and at the same time of year as the operational assessment.

Range Finding might also be described as a form of calibration. This is a process of determining the correct scoring for a set of benchmark papers and choosing a set of *anchor* papers. The anchor papers provide the operational definition of the scoring rubric and are a central component in the rater training process. It requires a significant sample of papers at each score point to provide a rich assortment of responses to support the thorough training of raters.

Table 5.6: Some Current Systems for Automated Scoring

AES Products	Method
Project Essay Grader (PEG)	**Regression / NLP***
Intelligent Essay Assessor (IEA)	**LSA** / NLP**
E-rater	**Regression / NLP**
IntelliMetric	**AI*** / NLP**
Bayesian Essay Test Scoring System (BETSY)	**Bayesian**
ETS I	Lexical-semantic
C-Rater	AI / NLP
Intelligent Essay Marking Systems	Pattern Indexing Neural Network
Automark	NLP
Schema Extract Analysis and Report	NLP
Paperless School free-text Marking Engine	NLP

*NLP: Natural Language Processing; **LSA: Latent Semantic Analysis; ***AI: Artificial Intelligence

Automated Essay and Constructed Response Scoring

In another context, Mason and Grove-Stephenson (2002) report that teachers in Great Britain spend 30% of their time grading papers: "… if we want to free up that 30% (worth 3 billion UK pounds per year to the taxpayer by the way) then we must find an effective way, that teachers will trust, to mark essays and short text responses." For large-scale assessment as well, the hand scoring of essays and other student work is expensive and time consuming.

Dikli (2006) and Valenti, Neri, & Cucchariarelli (2003) provide overviews of a number of possible approaches to automated scoring. Dikli provides a fairly detailed discussion of the first five scorers shown in boldface in Table 5.6, while Valenti et al. give more concise summaries of all in the table except *Intellimetric*. The first four listed have been applied to large numbers of essays in high stakes situations, although all seem continuously under development. For most high stakes tests, automated scoring is used in conjunction with a human scorer added to resolve inconsistent scores. The other systems are works-in-progress or are more research oriented.

Performance results are remarkably consistent across the systems, although the analyses report a variety of performance measures, i.e., correlations, percent exact or adjacent scores, making them difficult to compare. Many of the performance statistics were collected and reported by the developers and distributors of the systems. However, overall, the automated scorers appear quite comparable to human scorers.

While many of the developers protect proprietary aspects of their systems, the applications can be roughly grouped, with considerable overlap, into three general categories: statistical, artificial intelligence, and latent semantic analysis. As there is a growing emphasis on detailed, diagnostic feedback, most systems rely to a greater or lesser degree on the evolving science of *natural language processing* (NLP).

The first approach to automated scoring was statistical. In this category, *Project Essay Grader (PEG)* is by far the earliest entry (Page, 1966, 1994). PEG used easily observed approximations (*proxes*) for intrinsic measures (*trins*) of essay quality as predictors in a standard multiple regression analysis.

The dependent (or outcome) variable in the regression is the score assigned by a human reader. The process requires a set of calibration papers that have been scored to estimate the regression coefficients; these coefficients are then applied to unscored papers and a predicted score is obtained for each. The *trins* most mentioned are fluency, complexity, and articulation. The *proxes* could be total word count, word count by part of speech, and variation in word length, for example.

While this approach has proved effective and useful, it is often criticized for being statistical; that is, relying on correlational relationships with surface features and not inherent properties of the essay[6]. If the regression coefficients were made public, for example, it would conceivably be possible to subvert the system by just writing more words. The method is also limited in the type and quality of feedback that can be provided as an inherent part of the analysis. The count of the number of prepositions in the essay or the standard deviation of word length does not seem very useful for guiding the teacher or student.

Safeguards to detect anomalies, such as pairing with a single human rater, flagging unusual patterns of the *proxes*, and parsing with an NLP engine, are often incorporated into high stakes applications. More informative feedback can be derived from the NLP parsing along with the regression results. Even with safeguards in place, and enhanced diagnostics, the statistical approaches are less attractive than a method that uses NLP more directly to evaluate the students' work.

With subsequent advances in NLP, it became much more manageable to extract direct indices of essay quality. This is the direction that has been taken by PEG (Page, 1994) and by *E-rater* (Burstein, 2003; Burstein & Marcu, 2000). The new applications extract 100 or more linguistic and other features of the essay to use as the predictor variables in the regressions. This approach has been shown to be as effective as prediction from the proxy variables of the purely statistical, earlier models.

The statistical approaches are strengthened by the use of NLP to derive intrinsic indicators, but they still are limited by reliance on correlational relationships, not all of which are causal. The process requires the developers to imagine all reasonable features to describe the quality of an essay and to create algorithms to extract them[7]. The predictors are then derived by

[6] It can also be criticized for making up words like *proxes* and *trins*.

[7] This begs the question at what point does the extraction stop being a proxy and become an intrinsic.

stepwise regression or similar methods based on the correlation matrix and not on any linguistic or other theory.

Natural language processing is inherent in the *artificial intelligence* (AI) applications, like *IntelliMetric* (Elliot, 2003, Vantage Learning, 1999; Rudner, Garcia & Welch, 2006). Like statistical systems, the AI applications require substantial samples of human-scored papers to calibrate the engine. They use NLP to extract a multitude of descriptive indicators about each paper and arrange them to maximally differentiate papers that received different scores from the human readers. When processing unscored papers, the task is to determine which score category contains essays that most closely resemble the paper in question.

The AI approach probably mimics holistic scoring by human raters more closely than other methods. While this does not imply it is more reliable, it is probably more difficult to subvert. Proponents, in fact, argue that during calibration the engine *internalizes the collective wisdom of the human raters* (Elliot, 2003) with the intent of placing unscored essays in the same score categories as the human raters would. This can be viewed as either a strength or a weakness. In the current environment that places a premium on continuity with the existing measurement, it is a strength. On the other hand, AI does not eliminate the need for a trained pool of human raters and the scores cannot be better (i.e., more valid, reliable, or informative) than the human scoring would be.

Latent Semantic Analysis (LSA) (Lemaire & Dessus, 2001) is a theoretical strategy from cognitive psychology of the 1970s and uses some esoteric mathematics to build complex concept networks based on word-phrase associations. From a high level perspective, LSA analyzes a body of texts, constructs a very large matrix of word associations, and decomposes that matrix to construct meaning similarities among words and texts. It is possible to establish a connection between terms that did not occur together. The LSA networks tend to be highly correlated with the relationships established by humans.

Using LSA to score essays involves building the concept network for a domain, computing a *vector* for each essay in the high-dimensional space defined by the network, and choosing an existing vector that is similar to the one in question. The domain might be student-written essays used for calibration, similar to the statistical methods, or it might be relevant reference works for the topic addressed by the student essays. Scoring could be the holistic score assigned to the closest calibration paper by a human rater or it could be a continuous variable, such as the cosine[8] with the ideal response.

IEA, the best known implementation of LSA, has performed well in situations where students read a passage and write responses to prompts related to the passage. It seems capable of providing relevant, usable diagnostics in real time as well as providing a meaningful holistic score. It also has the potential to function without human scorers. When used in conjunction with humans, it can provide insights that are different from those made by the human readers.

Summary: Human versus Computer Scoring with Computer-based Assessment

1. The primary incentive for electronic testing is timely reporting of results. At this time, for high stakes, large-scale assessments, it probably is not practicable to provide instantaneous results. Receiving reports within a few days of testing seems more prudent and more manageable. This may change in the future or for lower stakes assessments. Alternatively, a number of states report preliminary MC results for teachers' use before the CR results and final scale scores are available.

[8] The result would not be reported as a cosine but would be converted into a more palatable scale score.

Like eTesting in general, automated essay scoring (AES) is also probably inevitable in some form and role. The existing models are promising, functioning with reliabilities similar to human scoring. There is also the potential for faster reporting of results and enhanced diagnostics. This may make it more appropriate and attractive for formative testing and instruction than for large-scale assessment and accountability systems.

2. For the immediate future, automatic scoring in high stakes situations will almost certainly be run in conjunction with human scoring. Most systems for AES will continue to require some amount of human scoring for calibration at the very least. The parallel systems with one human score and one computer-generated score are appropriate both to maintain public confidence and to safeguard against unforeseen anomalies the computer algorithms were not prepared to handle.[9]

3. The discussion heretofore has been concentrated on writing assessment and scoring of essays. In terms of AES, models for scoring writing are the most developed and the most used. Short-answer CR present greater challenges to the computer scoring engines. While it may seem counter intuitive that short responses are harder to score than long responses, it is actually a *sample-size* issue. A response with very few words does not adequately sample the student's proficiency to permit a computer algorithm to make an inference. The regression, AI, and LSA models require more information to function correctly.

[9] One frequently asked rhetorical question is, What do you do with the essay that is semantically and syntactically perfect but factually nonsense. Possible responses are, What should human scorers do with it? And, make the author a political speech writer?

4. Very short responses are perhaps best scored with pattern matching (Hirshman, Breck, Light, Burger & Ferro, 2000). These are quickly evolving beyond exact matches to more complex NLP processing applications. This now includes algorithms that can identify the appropriate information in a reading passage needed to answer and then evaluate whether the student's response agrees well enough to receive credit. The combination of constructing the key and evaluating the student's response make this approach very attractive for formative assessment, but perhaps not for large scale, high stakes.

FINAL NOTE: VALIDITY WITH AND WITHOUT CR

Issues surrounding validity deserve some additional comment. The decision about whether or not to include CR tasks often centers on what they add to the validity of the scores.

Face validity has been mentioned as an important aspect of the appeal of CR to educators and policy makers. CR tasks *look more like* what students need to be able to do. Presumably the purpose of, say, reading instruction is to equip students with the tools they need to decode, understand, apply, synthesize, assimilate, etc., more and more complex written text. The purpose of math instruction is to equip students to solve more and more complex problems. CR tasks can have a much more direct connection to these outcomes.

MC items tend to function at the more elemental level of the specific, distinct skills and tools needed to succeed at the higher level of solving problems. They assess whether the student can decode words or add numbers. They can be extremely useful in diagnosing what tools the student does or does not have. They have also been shown, over scores of years, to be effective for measuring achievement and proficiency. MC have been validated almost universally as useful predictors of the constructs educators are interested in affecting.

Because of the prevalence of MC, the constructs have tended to become operationally defined as the thing that the MC test measures. Including carefully crafted[10] CR tasks in the assessment should strengthen the overall case for validity. The construct being measured is independent of the item type to the extent that the result for the CR correlates well with the result from the MC. Consequently, the inferences that can be drawn are stronger and more general than if the assessment were based on a single item type.

The intent of education in all content areas is to equip students will the capabilities to perform more and more complex tasks. Writing is not logically different from other areas. However, MC items do seem even more contrived and less direct when applied to writing assessment. Constructed responses, specifically extended responses, have played a larger role and, as a result, writing assessment has been the leader in the development of hand and automated scoring. A separate window for writing is feasible, with some caveats. With electronic testing, some of the logistic and timing concerns will be mitigated.

[10] *Carefully crafted* goes without saying. Adding a poor item of any type never improves an assessment.

COMPONENT 5 – REFERENCES AND LINKS

Burstein, J. (2003). The e-rater scoring engine: Automated essay scoring with natural language processing. In: M. D. Shermis & J. Burstein (Eds.), *Automated Essay Scoring: A cross disciplinary perspective* (pp. 113-122). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Burstein, J. & Marcu, D. (2000) *Toward Using Text Summarization for Essay-Based Feedback,* Proc. TALN 2000 Conf., Lausanne, Switzerland, 2000. Retrieved Feb. 12, 2007 from http://citeseer.ist.psu.edu/burstein00toward.htm.

Dikli, S. (2006). An overview of automated scoring of essays. *Journal of Technology, Learning, and Assessment,* 5(1). Retrieved February 5, 2007, from http://www.jtla.org.

Elliot, S. (2003) Intellimetric: for here to validity. In Mark D. Shermis and Jill C. Burstein (Eds.). *Automated essay scoring: a cross disciplinary approach*. Mahway, NJ: Lawrence Erlbaum Assoc.

ETS Online Scoring Network: http://www.etsliteracy.com

Glas, C. & van der Linden, W. (Eds). (2003) *Computer-Adaptive Testing*, pp. 117–128. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Hirschman, L., Breck, E., Light, M., Burger, J.D., & Ferro, L. (2000). Automated grading of short-answer tests. In M.A. Hearst (Ed.), *The debate on automated essay grading*. IEEE Intelligent Systems, 27-31. Retrieved February 26, 2007, from http://pearsonkt.com/papers/IEEEdebate2000.pdf

IMOAT: The iCampus-MIT Online Assessment Tool. http://web.mit.edu/imoat/

Kukich, K. (2000). Beyond Automated Essay Scoring. In M.A. Hearst (Ed.), *The debate on automated essay grading*. IEEE Intelligent Systems, 27-31. Retrieved February 26, 2007, from http://pearsonkt.com/papers/IEEEdebate2000.pdf

Lemaire, B & Dessus, P. (2001). A System to Assess the Semantic Content of Student Essays . *Journal of Educational Computing Research,* 24-3, 305–320.

Mason, D. & Grove-Stephensen, I. (2002). Automated free text marking with paperless school. In: Proceedings of the 6th CAA(Computer Assisted Assessment) Conference, Loughborough: Loughborough University.

Page, E. B. (1966). Grading essays by computer: Progress report. *Notes from the 1966 Invitational Conference on Testing Problems*, 87–100.

Page, E.B. (1994). Computer grading of student prose, using modern concepts and software. *Journal of Experimental Education*, 62(2), 127–42.

Rudner, L. M., Garcia, V., & Welch, C. (2006). An Evaluation of the IntelliMetric^SM Essay Scoring System. *Journal of Technology, Learning, and Assessment*, 4(4). Retrieved February 26, 2007, from http://www.jtla.org

Valenti, S., Neri, F., & Cucchiarelli, A. (2003). An overview of current research on automated essay grading. *Journal of Information Technology Education*, 2, Retrieved February 5, 2007, from http://jite.org/documents/Vol2/v2p319-330-30.pdf.

Vantage Learning (1999) *Construct validity of IntelliMetric with International Assessment* (RB-323). Newtown, PA: Vantage Learning.

Component 6 – Hardware, Software, Staffing and Training Requirements at the State, District, and School Levels to Administer Statewide Computer-based or Computer-adaptive Assessments

This section examines the current state of technology-related infrastructure, staffing levels, and staff perceptions in South Carolina districts and schools. Understanding the technology requirements for system hardware, software, and bandwidth is absolutely critical to the successful implementation of an electronic testing program. The findings rely on information collected via a survey of district test coordinators, and technology staff. The findings serve as the foundation for making recommendations; identifying what, if any, technology readiness gap exists; what efforts would be required to close that gap prior to implementing an eTesting solution; and identifying possible training approaches.

This section begins with an overview of the types of electronic testing products available today, and identifies the specific technical requirements for several vendor offerings.

Test Delivery Mechanisms

Nearly all products used today for the electronic delivery of standardized large-scale assessments do so through one of two mechanisms: browser-based or smart client technology. The first mechanism utilizes web delivery of assessment content through common web browsers on the computer. Generally, the student starts a web browser such as Internet Explorer on the PC and logs on to a secure web site using a predefined username and password. From there the assessment content is delivered and student responses are captured through the browser. This approach is a fairly straightforward design, so the primary advantage to this approach is cost. In general, companies can deliver assessments through this approach at a lower cost because they are able to take advantage of browsers already installed in the schools and do not have to develop their own. However, this approach does come with some disadvantages, such as the following:

- Security—Generally speaking, this delivery system is less secure than the other commonly used delivery method, since it is not possible to completely prevent students from leaving the testing environment and potentially accessing other applications on the computer or the Internet. While some companies claim to have secured the browser by disallowing commonly used key sequences, a technically knowledgeable individual can circumvent these measures with little effort.
- Performance—Since the computers are connected to the Internet for the receipt of each test item and the transmission of the results, the performance or response time becomes much more dependent upon the Internet. Caching information, or queuing it in the PC's memory, may alleviate some performance issues, but also has its tradeoffs. There may also be some technologies at the school/district that might cause some performance issues, so some pre-work and testing is necessary. These technology considerations include firewalls, content filters, and pop-up blockers.
- Standardization of environment—Despite all efforts to make browser displays common, the type of browser used can still have an impact on how assessment content is displayed. This could potentially result in the same item being displayed slightly differently by two different browsers. For high stakes, large-scale assessment programs that demand standardization, this could introduce issues. Vendors will often mitigate this risk by only supporting a limited number of browser types or versions.

The second approach is through the use of smart client technology. This typically involves the testing vendor supplying the testing sites with a standard software package that must be installed on each machine used for testing. This package often includes a proprietary browser that is used only for testing purposes. Some variations of this approach also incorporate a local caching server to help avoid Internet performance issues. This smart client approach offsets the security and standardization issues of browser-based delivery by taking control of the computer desktop and operating system to set up a more structured and programmatically controlled environment. However, this approach has disadvantages as well, including:

- This software is typically more costly in terms of actual dollars because of its custom nature. In addition, it must be continually updated, tested, and distributed to keep up with rapidly changing technology environments.
- Because the nature of this software involves a footprint on each computer used for testing, there are vendor and local district/school staff hours required at each testing site to have the software installed.

The delivery approach chosen is dependent on the specific requirements of the assessment program. In general, the first approach (browser-based) is preferred for low stakes or formative testing programs where security and standardization needs are not as high. The smart client approach is often preferred in high stakes and/or large-scale programs with greater demands.

Vendor Analysis

In order to put South Carolina schools' current computer system characteristics into the context of the needs for electronic testing, the system requirements of five electronic testing delivery products have been analyzed. The list of vendors selected for analysis is not meant to be a comprehensive list of all vendors offering electronic testing products. It is intended only as a representative sample for purposes of determining an industry average set of system requirements. These vendors were selected based on their current usage within states for high stakes, summative test delivery, and the availability of accessing their recommended system requirements through public information sources. Should South Carolina proceed with electronic testing, system requirements and network configuration requirements for other vendors may become available through the Request for Information or Request for Proposal process.

The vendors selected for analysis within this report were grouped into three categories:

- Products typically used for **general student population testing.** This analysis includes four products in this group.

- Products that include features for the **accommodated student population.** This document provides information on three products that deliver tests enhanced with media, such as audio files, video files, or "read to me" features.

- Products that support **caching servers.** Two products are discussed in this group. The configuration approach includes additional computers that pre-load test content for the school or computer lab. This allows the testing computers to access the cache server for test content rather than taxing the Internet connection to retrieve content as students are testing. This architectural approach generally is targeted at improving the performance of the electronic testing experience.

The technology analysis within this section will focus on the configurations needed to test the general population, since both media-enhanced products and caching systems generally require a more robust computer profile than general population testing products. Although the focus is on the products for testing the general student populations, accommodations and caching server topics are addressed later in this section.

Most vendors provide <u>minimum</u> specifications needed to support the product as well as <u>recommended</u> specifications. Given that overall performance and user experience should be optimal when the <u>recommended</u> system requirements are met or exceeded, the <u>recommended</u> specifications were used in this analysis.

The following tables show the vendor systems used within the general population, accommodated population, and caching server portions of this analysis. Details of the recommended system requirements for each vendor/test delivery system are provided in the System Requirements Detail portion of Component 6.

Table 6.1

Vendors - General Population	Test Delivery Mechanism
Computerized Assessments and Learning	Smart Client
Internet Testing Systems	Browser-based
Pearson Educational Measurement Solutions	Smart Client
Vantage Learning	Browser-based
Vendors – Accommodated Population	**Test Delivery Mechanism**
Computerized Assessments and Learning	Smart Client
Measured Progress	Smart Client
Pearson Educational Measurement Solutions	Smart Client
Vendors – Caching Server	**Test Delivery Mechanism**
Computerized Assessments and Learning	Smart Client
Pearson Educational Measurement Solutions	Smart Client

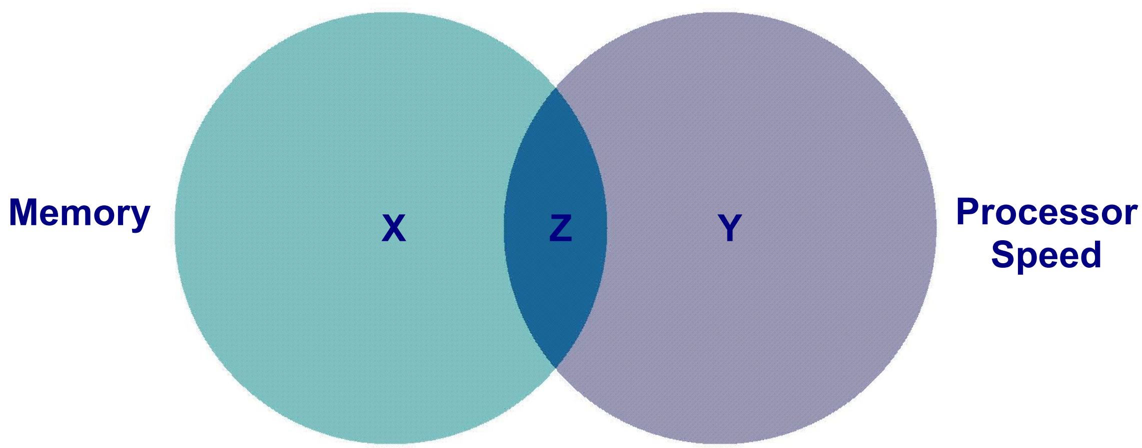
Analysis Method

After analyzing the test delivery system requirements, the system requirements were compared to survey responses to identify potential technology readiness gaps. A complete set of data that included every district and school was not available because 71% of the districts and schools responded to the survey. However, the robustness of the data that was collected allowed survey responses to be estimated for the non-responding districts and schools, and this adjusted set of data was used for this section of the analysis. For more details on the methodology used for creating the adjusted dataset, please see the Adjusted Dataset portion of Survey Methodology & Overview section of this report.

The process of analyzing the infrastructure data from the surveys is somewhat complex, due to the number of system components that must be reviewed to determine the technology gap. Additionally, since each of the vendor products has its own set of technology requirements, this adds another dimension to the analysis. The survey gathered information on several technology components, such as memory, processor speed, operating system, browser, and type of Internet connection. Although the data provided a clear aggregated picture of the current state of South Carolina's technology readiness in each area, a clear individual computer profile is not available. However, there is enough data to form conclusions and recommendations.

Figure 6.1 provides some background for the analysis of the technical components. Imagine only two system requirements must be met: memory and processor speed. The survey provides the percentage of computers meeting certain ranges of memory sizes and percentages of computers with certain processor speeds, as shown by the two circles in the diagram. Only the computers that meet both requirements would fall within the intersection in this figure.

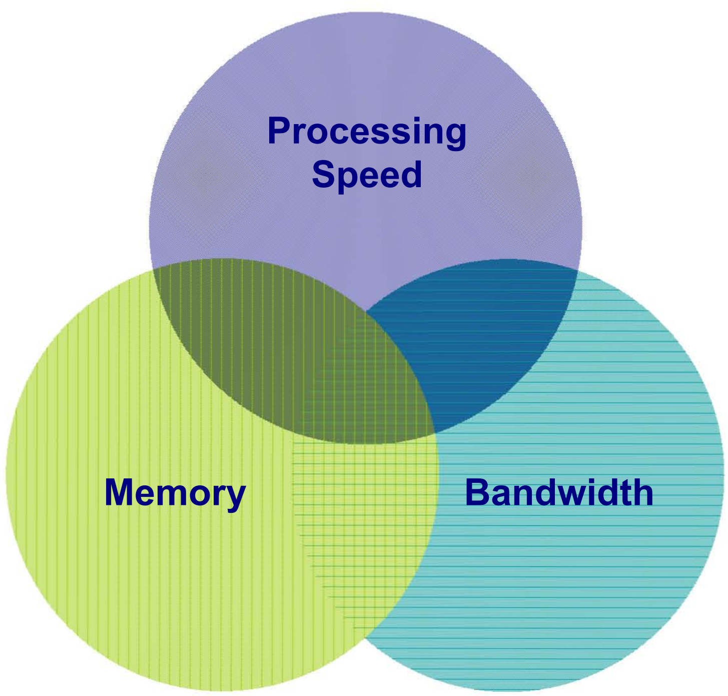
Figure 6.1: Overlap of Memory and Processor Speed



Therefore, it is possible for some of a school's computers to meet or exceed memory requirements and some computers to meet or exceed processing speed requirements. Knowing that X% of computers meet the memory requirement and Y% meet the processor requirement is not sufficient information to determine the number of computers (Z) meeting both requirements. It is not clear whether the computers that meet the memory requirements are the same computers that meet the processor requirements, nor should it be assumed that they would be the same.

Since the system requirements include a range of components (such as memory, processor speed, operating systems, browsers, and screen resolution), the greater the number of variables introduced, the more likely that any single computer will fail to meet at least one requirement. In general, the requirements that have the greatest impact on performance during testing or the selection of a testing vendor are bandwidth, processing speed, and system memory[1].

Figure 6.2: Overlap of Three System Requirements



[1] Other factors, especially operating systems, are vital, but these three specifications are used to simplify this example.

Given the gap in data describing a workstation profile, it is still possible to estimate an upper and lower boundary for the percentage of computers meeting all three critical requirements: memory, processor speed, and bandwidth on a per-product basis. This provides a best-case and worse-case scenario for South Carolina's technology readiness gap based on the four general population testing products analyzed. The best case assumes the greatest overlap in computers meeting each requirement collectively. The worst case distributes the "failed" requirements across the greatest possible number of computers. For example:

Table 6.2

	Vantage Learning System Requirement	% of computers meeting requirement	% of computers failing requirement
Bandwidth	Broadband or better	56.45%[2]	43.55%
Processing Speed	1 GHz	64.24%	35.76%
Memory (RAM)	256 MB RAM	86.30%	13.70%

In the best case, hypothetically, the 56.45% of South Carolina's computers that meet the bandwidth requirement also meet processing speed and memory requirements. The best case is calculated as the minimum of the percentages meeting the requirement; no more than 56.45% of the South Carolina computers will meet the requirements of the Vantage Learning system.

In the worst case, again hypothetically, the computers that fail the bandwidth requirement are different computers than those that fail the memory requirement, and are different computers than those that fail the processing speed requirement. The worst case is calculated as the sum of the percentages of failure to meet the requirement; up to 93.01% of the South Carolina

[2] This excludes computers with dial-up and those responses of "other."

computers may not meet the requirements of the Vantage Learning system, which could mean that as few as 6.99% may satisfy all three critical system requirements.

The following table shows the best case and worst case for each of the products analyzed. A per-system breakdown of the percentages, as shown above for the Vantage Learning system, is available in the System Requirements Detail section of Component 6.

Table 6.3

	Best-Case Scenario	Worst-Case Scenario
Computerized Assessments and Learning	56.45%	53.06%
Internet Testing Systems	56.45%	40.10%
Pearson Educational Measurement Solutions	56.45%	28.95%
Vantage Learning	56.45%	6.99%

The above table only considers three system requirements:

- bandwidth,
- processor speed, and
- memory (RAM).

Adding additional system requirements may reduce the best-case scenario and are guaranteed to increase the severity of the worst-case scenario. In addition, the bandwidth requirement percentages above do not consider 42.09% of the survey responses because it cannot be assumed the response "Other" was selected in cases where the respondent did not know the connection type or whether Internet connections to the computer were present. Bandwidth is a critical requirement and will be discussed at length in the bandwidth subsection.

The remainder of this section analyzes the survey results from the adjusted dataset using mean percentages and comparing results to the system requirements of the vendor products. The results of this analysis are presented separately based upon the specific system requirements.

Current Technology in South Carolina

This section presents the survey results and provides an analysis of the results compared to the vendor requirements for the general population testing products. These components will be addressed individually:

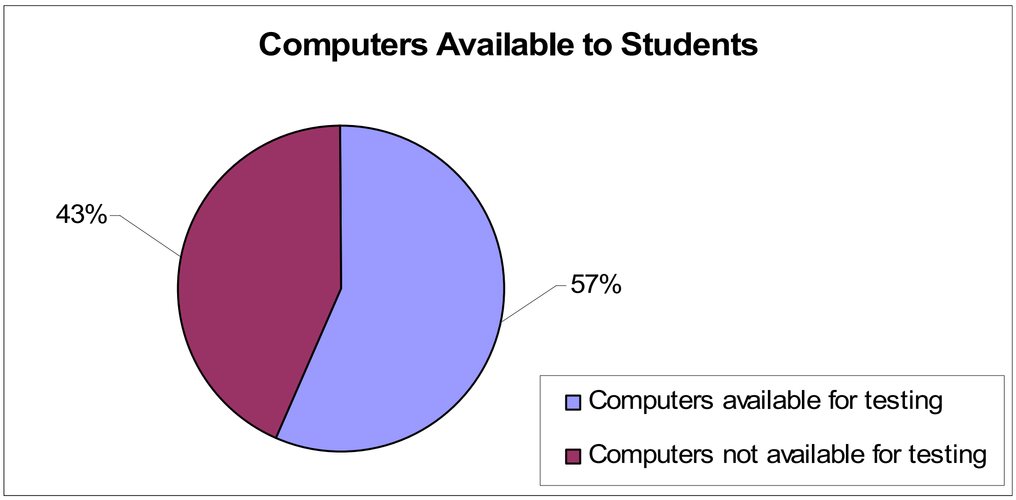
- Computers available to students
- Processing speed
- Operating system
- Memory (RAM)
- Display
- Filters / Firewalls
- Browser Compatibility

A summary of the technology gap information will follow the analysis of these individual components.

COMPUTERS AVAILABLE TO STUDENTS

Statewide, a total of 177,604 computers are available to students within schools and a maximum of 100,372 computers could be made available for testing. An overwhelming majority (93%) of survey respondents indicated students would complete standardized, computer-based assessments in a computer lab environment.

Figure 6.3

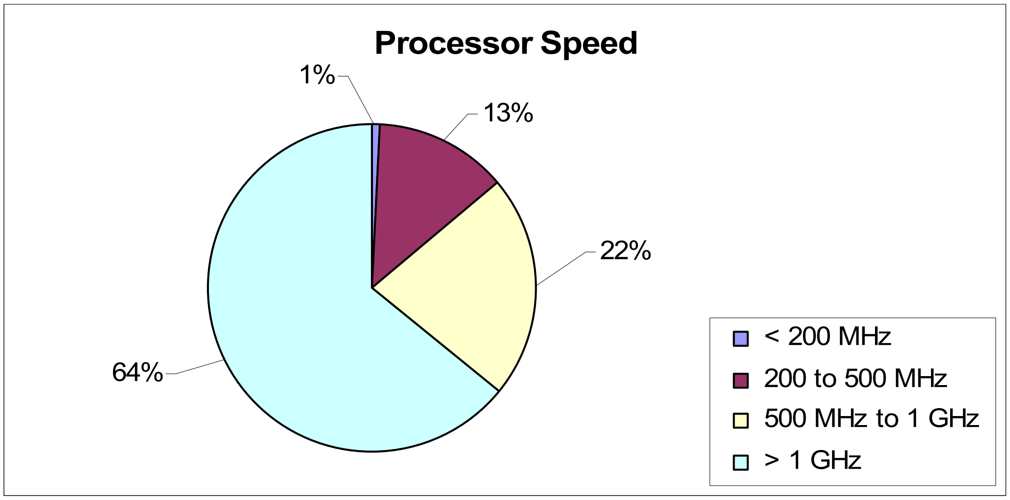


The survey responses indicated that 43% of the computers available to students would not be available for testing purposes. The survey did not ask why such a high number of computers (77,232) would not be available. If these computers could be made available for testing, either by temporarily relocating the computers or administering tests in a non-lab environment, it is likely that fewer computers would need to be purchased, as indicated in Components 7 & 9 of this study. In order to determine if these computers could be used for testing, additional conversations with district and school personnel would be required.

The percentages used in the following analysis are based upon only the computers available for testing.

PROCESSING SPEED

Figure 6.4

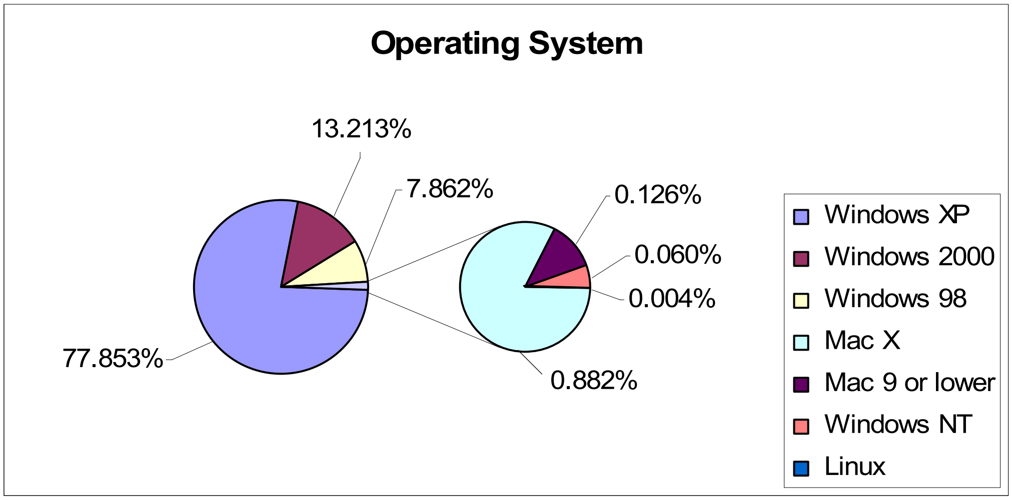


A majority of the computers available for testing (64%) currently meet the processing requirements for all four testing products. The remaining 36% of the computers would support the requirements for two and potentially three testing products. Two of the product's requirements fall within the 200 to 500 MHz range, and the actual number of computers meeting the requirements within that range cannot be assumed. However, an upgrade to 1 GHz or better would increase the state's options for vendor selection, as well as provide a better user experience in general.

Approximately 14% of the computers have 500 MHz processors or slower. These computers could support only one vendor's product, significantly reducing the state's options for test vendor selection. These processors should be upgraded in order to support electronic testing.

OPERATING SYSTEM

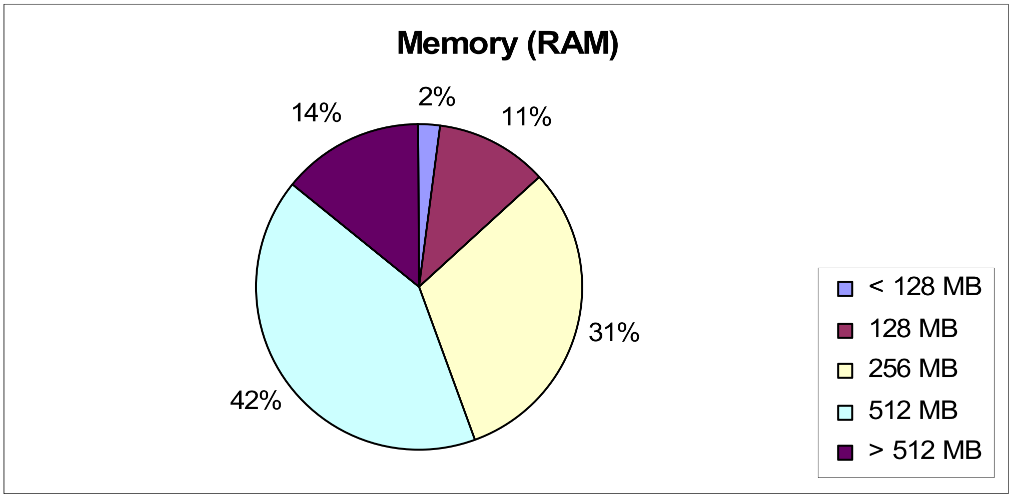
Figure 6.5



The majority of computers available for testing (78%) currently run Windows XP, which is a compatible operating system for all testing products. Of the remaining computers, 7.8% are currently running on Windows 98, which is supported by two of the products. However, Microsoft has discontinued support of Windows 98 effective July 11, 2006, and testing vendors have indicated their products will not support the Windows 98 operating system after this year. Microsoft will support Windows 2000, which is on 13% of the computers currently through 2010, so the computers currently using that operating system may need new operating systems at that time, depending on vendor specifications. Machines running Windows 98, Windows NT, Mac 9 or lower, or Linux (a total of 8% of the computers) would need an operating system upgrade in order to support electronic testing.

MEMORY

Figure 6.6



Only 2% of the computers available for testing would be unable to support any of the testing products, as they currently have less than 128 MB RAM. Currently, 11% of the computers have 128 MB RAM. This meets the memory requirement for only one of the testing products. Approximately 86% of the computers available for testing met the RAM requirements for three of the products reviewed.

In order to reach the 256 MB memory recommendation, which satisfies all of the testing product specifications, 13% of all computers would need an upgrade.

DISPLAY

The majority of the computers in South Carolina currently have 15 or 17-inch monitors. These screen sizes would be compatible with all products reviewed. Most systems require either a minimum 800 x 600 pixel screen resolution or a 1024 x 768 pixel resolution. The higher resolution reduces the scrolling required during testing and presents the student with a better user experience.

Screen size is not the only factor that affects display and resolution. The video cards within the computer may affect the computer's ability to display the higher resolution. Display may not be a concern for newer machines, but may affect older computers. The average number of computers less than 18 months old per school was 39. Schools tend to have the highest number of computers between 19 and 48 months old. Looking at the average percentages, 47% of computers within a school are between 19 and 48 months old.

Given this information, some computers may need video card upgrades or new monitors. However, it is not possible to accurately estimate that percentage. Based upon industry trends over the past 48 months, it is likely that the current computers will support the higher resolution display. Because the drawback to a computer that does not support the higher resolution is scrolling during test administration, this specification may not be a factor in the technology gap.

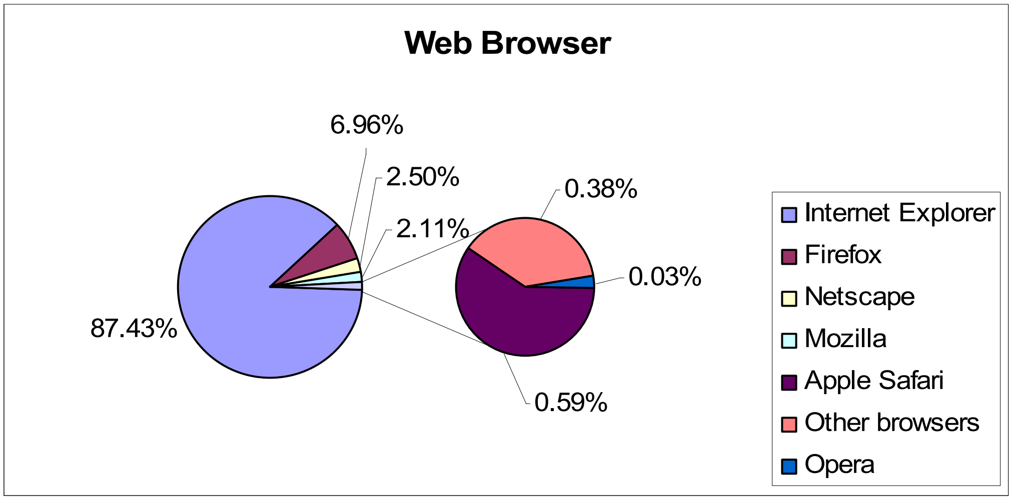
FILTERING/FIREWALLS

The majority of schools (87%) currently allow the industry standard HTTP and HTTPS connections to computers. For the computers that do not currently allow the connections, a network administrator can change this setting easily. Firewall products are utilized in 90% of the schools. Regardless of the vendor product selection, connections to certain external, vendor-specified addresses would need to be allowed. Again, this is a simple configuration step.

As schools and districts implement software safeguards, such as content filters or pop-up blockers, there will need to be a testing activity to ensure that the vendor's electronic testing solution can peacefully coexist with these safeguards.

BROWSER COMPATIBILITY

Figure 6.7



Approximately 13% of the computers available for testing may need a different browser in order to administer web-based tests. The majority of the state's computers (87%) use Internet Explorer, which is compatible with three of the four products. The other system did not specify a browser compatibility requirement.

TECHNOLOGY GAP BREAKDOWN

The following is a summary of the technology gap, excluding bandwidth, which exists for general student population testing. Bandwidth will be discussed separately in a following section. Bandwidth is a key factor to consider in terms of the infrastructure, and is somewhat more complex to analyze. The caching technology approach described at the beginning of this section can mitigate some of the bandwidth issues, and this will be discussed later in this section also.

The following is a summary of the analysis of the technology components discussed so far as they apply to computers which would be available for testing:

- 36% of computers will require a processor upgrade depending upon vendor selection.
  - 14% of computers can only support one vendor's product and should have a processor upgrade in order to support electronic testing.
- 9% of computers require an updated operating system.
- 13% of computers require a RAM upgrade to reach the 256 MB RAM requirement
- 13% may need a browser upgrade or change

The possible costs associated with closing this gap are addressed in Components 7 & 9 of this study.

INTERNET CONNECTION/BANDWIDTH

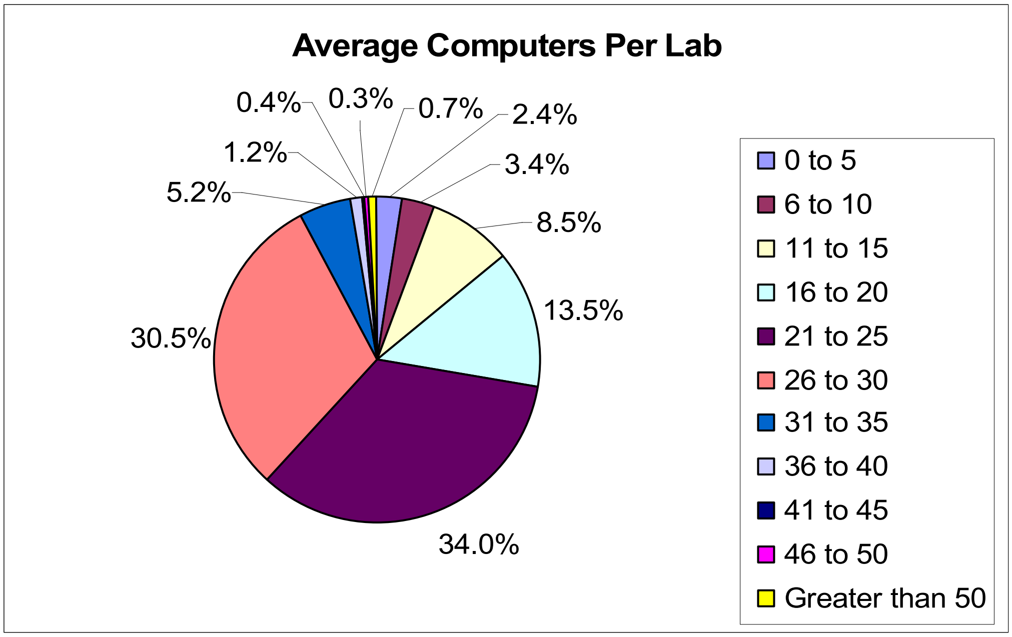
Bandwidth, specifically the bandwidth at the testing location, is a key technology factor when considering the implementation of an electronic testing program and is likely the most common problem faced when administering electronic tests. Bandwidth refers to the amount of data that can flow into and out of a particular location in a given time period. The amount of bandwidth needed is dependent upon the total number of students taking the test, or accessing the test information, within a specific time period. This is referred to as the number of concurrent users, or the number of students taking the test at the same time.

Bandwidth requirements also vary dramatically from one testing vendor to the next. In the state of Virginia, a smart client test delivery approach is used and the system calls for a minimum of 6 Kilobits per second/per user. System requirements specified by Kansas and Idaho show that a school with a T1 Internet connection can support up to 96 concurrent users, or approximately 15 Kilobits per second/per user. Each state must work closely with its testing vendor to determine the most effective approach for assessing bandwidth requirements.

Accurately measuring the bandwidth availability at the testing locations presents a unique challenge for school technology coordinators and testing vendors. Many schools lack the tools to accurately determine their current bandwidth capacities or usage, although some vendors have tools that will provide a snapshot of the bandwidth capabilities (at a moment in time). Also, bandwidth usage can vary dramatically depending on what users are doing from one hour or day to the next. This usage is not limited to student or lab usage, but may include other network traffic, such as eLearning activity, teacher activity, or administrative staff utilizing the connection. Many states including Virginia, North Carolina, and Kansas have worked with their vendors to develop special tools that schools can run to assist in determining bandwidth capacity at the testing sites.

When determining the bandwidth requirements baseline, schools and districts must consider the total number of students accessing the Internet to load test information during a testing session. This number must be multiplied by the vendor's recommended bandwidth requirements. As previously mentioned, one vendor recommends using 6 Kilobits/second per student as a rule of thumb for creating this estimate. Another recommends approximately 15 Kilobits/second per student for estimation purposes. For the following discussion, the rule of thumb used is 10 Kilobits/second per student.

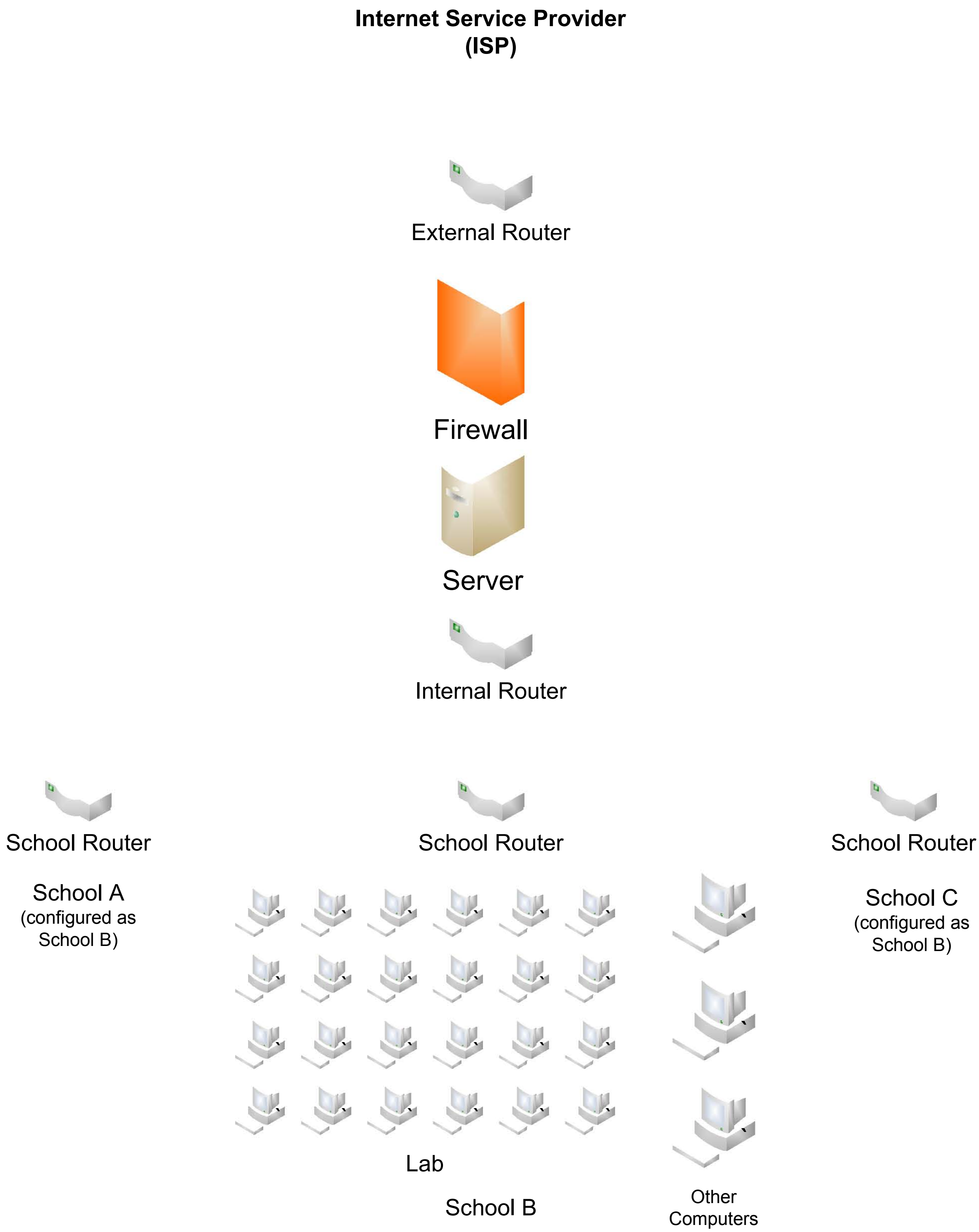
Figure 6.8



The average computer lab in the state has between 21 and 30 computers per lab, with the state mean being 24 computers per lab. This would indicate that the average school with one lab would need, at minimum, a 240 Kilobits/second connection.

Additionally, the infrastructure may be stressed by an entire school district sharing a single Internet connection. The staff at the schools may not know if multiple schools share the Internet connection, or if they have a dedicated connection. This could greatly affect the network requirements needed to support electronic testing. Therefore, the minimum bandwidth required from the district office to the Internet is the sum of the bandwidth needed for each school. The following figure illustrates the requirement for a district with three identical schools. Using the example above of the school with one lab and the following illustration, a district with ten such schools would require a connection with a minimum of 2,400 Kilobits/second.

Figure 6.9


Internet Service Provider
(ISP)
External Router
Firewall
Server
Internal Router
School Router
School A
(configured as
School B)
School Router
Lab
School B
Other
Computers
School Router
School C
(configured as
School B)

In general, T1 lines are rated to deliver 1,544 Kilobits/second and T3 lines are rated at 45,000 Kilobits/second. The actual performance of these connections is slightly less due to overhead. Therefore, a T1 line delivers approximately 1,200 Kilobits/second and a T3 line delivers approximately 42,000 Kilobits/second. In Figure 6.9, a T1 line would support only half of the network traffic. Two T1 connections could serve the district and support testing efforts. A T3 connection would be an alternative, but would likely be too expensive for the size and traffic used by the district.

Based on this information and survey data the following findings can be made:

- 4 % of the computers use dialup or a cable modem and may be better served by upgrading their bandwidth in order to mitigate performance issues during testing.
- 20% of the computers utilize a T1 line and could support approximately 120 students testing concurrently per T1 line[3].
- 33% of the computers utilize a T3 line or greater and could support approximately 4200 students testing concurrently per T3 line[3].
- 42% responded with "Other" as the type of connection used. It is possible that survey responders that were unaware of the actual Internet connection type simply selected "Other" because an "Unknown" response was not available. It is also possible that this response was selected when there is no Internet connection.

In addition, 3.6% of the computers connect to the Internet via a cable modem. These connections could be faster or slower than T1 lines, depending upon the provider. Some services advertise residential services of up to 8 Megabits per second, or 8000 Kilobits per second. DSL

[3] Assuming little or no other online activity is occurring during testing time.

and cable may also offer different bandwidth "upstream" (computer to ISP) as offered "downstream" (ISP to computer) which could affect performance.

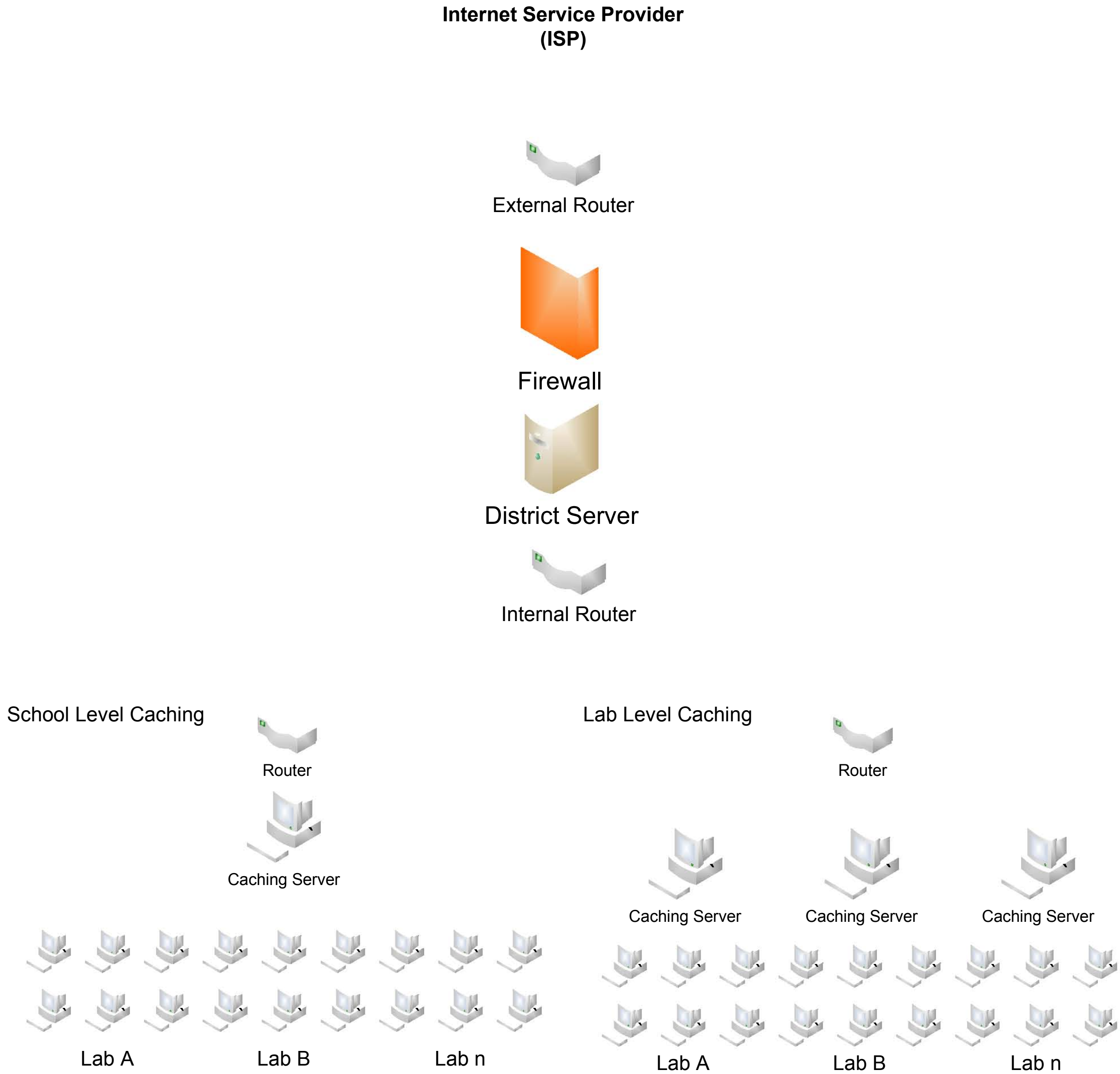
Caching Server Requirements

To mitigate performance issues sometimes associated with computerized testing, two[4] of the vendors researched offer a caching system that allows test content to be preloaded at the school or computer lab levels. This allows the testing computers to access the cache server for test content rather than taxing the Internet connection as students are testing. Generally, the system requirements for the caching server are greater than the requirements for student workstations. Computers that would meet these requirements may not be available in schools or districts currently.

A cache may reduce the bandwidth requirement. However, doing so also makes the network configuration more complex. One vendor has been able to utilize a lab-level caching server in a low capacity network environment to achieve an acceptable performance level for a ratio of 200 student workstations to one cache server. The same vendor has been able to utilize a school-level caching server in a high capacity network environment to successfully test all students within the school from the single cache server. The optimum network configuration would be a factor in how many machines meeting the caching server requirements are needed.

[4] The Measured Progress system, which administers the CATS Online assessment in Kentucky, offers a type of caching, but has been excluded from this comparison due to the accommodated nature of the CATS test and its more robust system requirements.

Figure 6.10


Internet Service Provider
(ISP)
External Router
Firewall
District Server
Internal Router
School Level Caching
Router
Caching Server
Lab A
Lab B
Lab n
Lab Level Caching
Router
Caching Server
Caching Server
Caching Server
Lab A
Lab B
Lab n

According to the survey responses:

- 86% of the computers available for testing meet the lower of the memory requirements (256 MB RAM) for one vendor's lab-level caching server.
- 55% meet or exceed the higher 512 MB memory requirement for school level caching.
- No more than 14% meet the 1 GB memory requirement for the other vendor system.
- 86% meet the lab-level server processing speed requirement.
- 64% meet the 1 GHz or higher processing requirement.
- 78% of the computers are running on the Windows XP operating system, which is a supported operating system for both vendors in either lab or school level configurations.

Although it seems that these computers may be available within schools currently, this would reduce the number of student workstations available for student testing. However, based upon the specifications, it appears that the test administrator could utilize these computers during the test for management functions as needed.

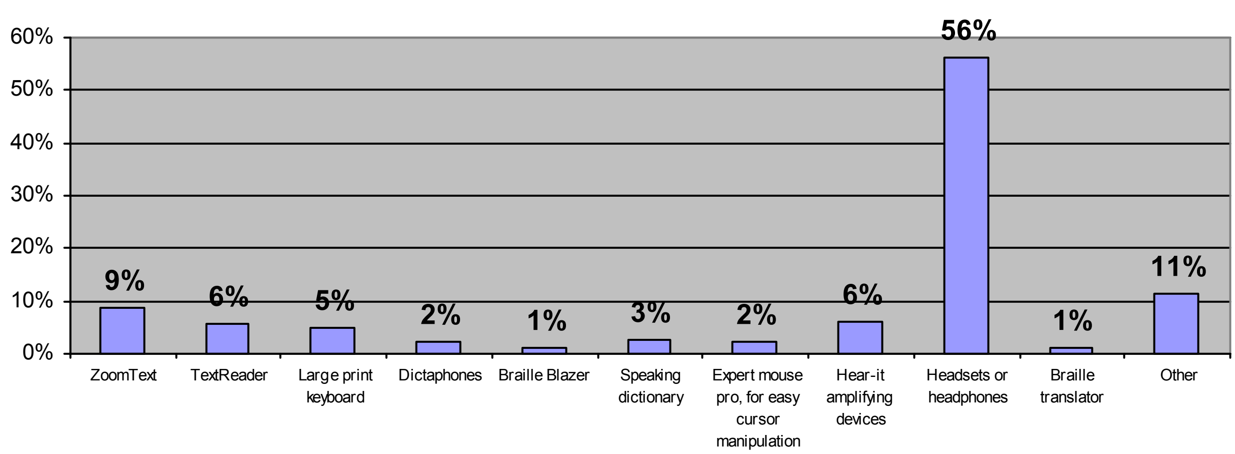
ACCOMMODATIONS

Three of the vendors researched offer test delivery enhanced with media. This may include a text-to-speech function, audio files, video files, or Flash-based content. These capabilities may be leveraged for those students requiring accommodations, such as those requiring read-alouds, or as a more robust delivery mechanism for the general population. Video or Flash-based content may allow for more complex item types, too. Although enhanced tests may be administered to the general population, the system requirements for this type of test delivery are slightly higher than those for general population testing.

The greatest impact to system requirements is in bandwidth, processing speed, and memory. Of the three products analyzed, two products require a Pentium IV processor, which is at least 1.5 GHz. Based on the survey results, 64% of the computers available for testing have a processor speed greater than 1 GHz. However, it cannot be assumed that all of these computers meet the 1.5 GHz requirement. The RAM requirement for two products remains the same, or increases to 256 MB, a level already discussed (86% of the computers have 256 MB RAM or greater). The third product has a sharp increase in the memory requirement to 512 MB RAM; however, 55% of the computers currently meet or exceed this higher memory requirement.

Aside from the processing requirements, additional hardware or accessories may be required, depending on the students' needs. When delivering a test that utilizes text-to-speech technology, speakers, or more likely headsets, would be required. South Carolina schools already have certain accessories in place, the most common being headsets or speakers, as shown in Figure 6.11.

Figure 6.11



The addition of media to the test administration may also have a significant impact upon the bandwidth or Internet connection requirements. However, all three products offer, recommend, or even require some type of caching system. Two of the caching system requirements are reviewed within the Caching Server Requirements section. The third system was excluded from the caching analysis because the system requirements were specific to an accommodated administration, Kentucky's CATS Online assessment, which is administered only to special populations. The caching system requirements for this system were by far the highest of the systems reviewed and few, if any, current computers in South Carolina schools could meet those requirements.

Northwest Evaluation Association (NWEA) — Measures of Academic Progress (MAP)

The analyses throughout this component focus on test-delivery systems currently being used by states, including South Carolina, in a high-stakes, summative testing environment. However, NWEA's MAP system requirements were also analyzed due to MAP's widespread usage in South Carolina for formative/diagnostic purposes. Details of the recommended system requirements are provided in the System Requirements Detail portion of Component 6. Based on this analysis, nearly every computer available for testing could support the MAP testing system requirements for student workstations.

The NWEA product was not included in the detailed analyses of test-delivery systems currently being used for high-stakes, statewide testing or the best-case/worst-case technology gap analysis, as NWEA's assessment system is currently used only for lower-stakes, formative/diagnostic testing. Nor did this study specifically compare products based on low-stakes, formative versus high-stakes, summative test situations, although most of the products researched could be used for the delivery of either type of assessment.

Emerging Technology

This section reviews the potential for using new or emerging technologies in the assessment program. This may include use of PDAs, tablet PCs, handheld input devices, or other types of hardware. Many of these devices have had limited application in assessment programs to date. However, as assessment programs continue to evolve, so will the use of technology. Two key issues to consider with this new type of hardware are:

- the standardization of the testing environment and
- the applicability of these devices to learning activities.

As discussed earlier, large-scale standardized assessments, such as those given in South Carolina, often require a high degree of environmental standardization to ensure equity across the state. Introducing new test delivery mechanisms requires great care and research through comparability studies to maintain that level of equity. Also as discussed in other sections of the study, there is a strong beneficial correlation when students are instructed using the same or similar technology to the assessment program. Therefore, the introduction of emerging technologies into the classroom setting will provide the greatest value when the technology can be utilized for both instruction and assessment.

Current research on handheld technology in classrooms—compiled in *Ubiquitous Computing in Education*: *Invisible Technology, Visible Impact* (van 't Hooft & Swan, 2007)—suggests that handheld computing devices have indeed gained some traction in the K-12 educational arena. The research indicates that student engagement in learning and academic achievement can increase with the appropriate integration of handheld technology into classroom instruction, student research/data collection, student self-monitoring of progress, and classroom assessment.

However, most of the research regarding handhelds and assessment is in the area of formative/diagnostic/informal testing. This may include teachers using handhelds to record observations (e.g., DIBELS early literacy assessments offered via handhelds by Wireless Generation's mClass:Dibels) or students using handhelds to respond to a question the teacher has on a classroom screen, whereby responses are anonymous and immediately aggregated for the teacher. The advantage seems to be the immediacy of results that are accessible through easy-to-use data-management tools in the classroom.

Currently, no large-scale, high-stakes NCLB or end-of-course high school assessments are being administered via handhelds, as issues of screen size, student proficiency with the devices, small graphic displays, miniature keyboards, and the overall difference in presentation via handhelds vs. paper/pencil raise serious concerns relating to comparability, as well as construct validity.

Additionally, one handheld study noted that older students commented, "that it was difficult to take notes because the onscreen keyboard was too small and inconvenient for inputting letters, and the handwriting-recognition software would not allow them to take notes fast enough." Younger students "often had issues with text input due to their limited development of fine motor skills." (Namsoo, Norris & Soloway, 2007, p. 35)

When contemplating the possibility of employing handheld technology for large-scale assessment, it may also be useful to consider the observations of Sarah Susbury (Director of Assessment, Virginia) at the Expert Panel Meeting held in Columbia, SC, on March 28. Ms. Susbury noted that having a standard environment in schools (including having "the same real estate on the screen"), as well as considerations of the work environment students enter once they leave school, led Virginia away from the use of handhelds for statewide assessments.

The XO Laptop, sometimes referred to as the $100 Laptop, is another so-called emerging technology that has been discussed as a possible device for use in schools. Until recently, the One Laptop Per Child (OLPC) non-profit organization has focused its efforts solely on distribution to developing nations. However, in April 2007 Nicholas Negroponte, head of the OLPC, announced that in response to interest from 19 state governors, OLPC will consider selling the laptops to governments in the United States. Negroponte said the computers will enter mass production in September if the OLPC received orders for at least 3 million devices. He noted that U.S. schools could receive the laptops by the end of the year, but at a higher, unidentified price "because more resources are invested in American education than in developing nations, even in the poorest U.S. regions."

Although the published system specifications of the XO laptop may support two of the testing products analyzed, it is not clear whether this will be a viable solution in the immediate future. First, the device does not have a hard drive, which significantly affects the type of test delivery mechanism used in electronic test delivery. Second, the original specifications for the XO include a Linux operating system, which may not be supported by the testing vendors. Although OLPC has announced they will configure the laptops for use with the Windows operating system, the licensing of the operating system will likely increase the cost of the device. The Linux operating system may also present new support challenges for school technical staff because it is an open-source product and does not have a specific company behind it to provide technical support. If the XO is to be considered as a solution for South Carolina, it should be thoroughly researched by both the state and the selected electronic testing vendor.

Staffing & Training

While understanding the technology is key to a successful implementation, it is equally important to study and address the human element. Moving from a paper/pencil to computer-based assessment would represent a significant process change for nearly all those involved in the testing program. From content specialists who must be concerned with how students will interact with items on the computer, to test administrators who must address student concerns during testing, and school or district technology coordinators who are responsible for network availability and perhaps scheduling of computer time, there will be many new processes that will need to be explained and new training programs initiated.

The survey allowed three opportunities each for district-level and school-level respondents to provide comments regarding electronic testing. Although the comments spanned many areas from security to power supply to “we’re ready now,” two themes specifically related to staffing concerns emerged. The first involves the length of the testing window given the computers available and the additional staff needed to manage students within classrooms and labs. The current level of staff available for administering an electronic test is addressed in Component 8 of this report. Recommendations for computer availability and infrastructure are addressed in Components 7 & 9 of this report.

The second theme was the availability of technical staff. It seems that the perception among South Carolina staff is that any issue that arises during testing would require a staff member with technical expertise to correct. This is a valid concern, as the tension of a new system and the importance of the test results magnify the perception of any technical glitches that may arise. This perception could be attributed to the previously mentioned change in process for testing, a natural aversion to change, and recent media coverage regarding poor system performance during testing.

Experience in other states has shown that successful training and professional development programs are vital to establishing support from key stakeholders throughout the state. Developing high quality training models begins with establishing the needs of the stakeholders. In the current paper/pencil environment, the needs of these stakeholders are typically well-known based on years of experience and refinement of course models. The introduction of electronic testing introduces new stakeholders to the process and often new approaches for the delivery of training.

The partnership between the testing vendor and the state is vital to creating a training program that addresses staff concerns while instilling technical confidence in the test administrators. Although the technical requirements and initial configuration of the workstations may require staff with technical knowledge, a person with average computer skills can usually handle many of the alerts or error messages. Each of the researched vendors offers comprehensive user manuals, practice tests, and tutorials that would help the key stakeholders become more familiar and comfortable with the system before test day arrives. Working with the vendor to provide each testing location with documents, such as "Frequently Asked Questions," documentation outlining the resolution of errors, and an escalation plan (i.e., when district technical staff versus vendor customer support should be contacted to correct a problem) will assist in staff comfort.

Another vital partnership is between the technical departments within the state system and the test coordinators. This partnership can be established by providing training for both groups together, which allows the staff to hear the concerns from their counterparts to help formulate a plan to handle issues as they arise.

When Idaho moved its Idaho Standards and Achievement Tests (ISAT) from paper/pencil to computer-based, it conducted visits to districts to discuss the purpose of the change and to get feedback, which was then incorporated into the formal training of the users. They also developed specific training courses for technology coordinators in the districts in addition to the DTC training, as this was a key group of stakeholders who were not part of the assessment process previously. At the Expert Panel discussion, individuals closely associated with assessment programs in Indiana and Oklahoma stressed the need for face-to-face training sessions with test coordinators in their respective states during the conversion process. They emphasized the best way to reduce fears is through "over communicating" throughout the process, providing opportunities for field personnel to voice concerns and having those concerns addressed. The state of North Carolina used mobile testing labs as part of their training process to simulate all phases of an actual testing environment. This greatly enhanced test administrator's comfort level with this system, as there were fewer surprises on the day of testing. Another observation from the Expert Panel discussion is that although vendors understand their system, they will not be as adept at handling questions related to state policy. State level staff will be needed to address policy concerns throughout training efforts.

Based on these experiences, a training plan should include the following:

- **Communicate early and often:** Discuss the goals of moving towards electronic testing and the implementation plan, and listen to staff feedback. Encourage district technical personnel and test coordinators to work together, and include both groups in training efforts.

- **Demonstrate the system as much as possible:** Talking about how the system will work, seeing how it works, and actually using it are very different things. Allow state representatives, technical staff, and district/school staff the opportunity to simulate a test environment. The more exposure staff has to the test system prior to test day, the higher their comfort level.

- **Work with the testing vendor to get the most out of practice tests, tutorials, and user manuals**, as well as create plans or processes for how to handle alerts or error messages on student machines on test day. If possible, create simulations that include alerts that staff must resolve during simulations.

Regardless of the training program implemented, communication is a key aspect to establishing support from the key stakeholders throughout the state.

Summary

Although more than 177,000 computers are available to students within the state, 43% of those computers would not be available for testing. Among the computers available for testing and the technical specifications of the testing products analyzed, the greatest barrier to testing (the system requirement with the highest "failure" percentage) is processor speed: 36% of the computers would need a processor upgrade, and 14% of the computers could only support one of the testing products analyzed. The second major barrier is system memory: 13% of the computers need an upgrade to reach 256 MB RAM system requirement. A majority of the computers currently run a supported operating system (91%) and supported browsers (88%). Based upon the average age of the computers, many of the computers that would require processor or memory upgrades may already be scheduled for replacement.

Most of the schools utilize content filtering and firewalls. These products will still allow testing, but would need specific configuration changes to authorize certain Internet addresses during testing. Other system configurations may need to be adjusted during testing. This may include changing pop-up blocking, allowing session cookies, or enabling or disabling applications that run automatically (e.g., email notifications or virus scanning software).

A key area within the technology infrastructure is how many students will be able to test simultaneously given the current bandwidth at the school. Although the bandwidth needs may be reduced by utilizing a cache for test delivery, the number of students supported by each Internet connection may not be adequate to support the levels of concurrent users anticipated. The data gathered within the survey does not include whether multiple buildings share Internet connections, or if there are dedicated lines per school. This will have a great impact on performance during testing. Approximately 24% of the computers available for testing could support 120 students testing simultaneously per T1 line, assuming little or no other online activity during assessment time.

Three of the vendors analyzed have products that include multimedia test delivery. This includes, at minimum, a text-to-speech feature that may be utilized for students who require a read-aloud accommodation. These products may require a more robust system configuration or the implementation of a caching server to deliver the test content to all students within acceptable performance levels.

The staff members who responded to the survey provided general comments regarding electronic testing. Based on these comments, staff members are concerned with the number of computers available for testing in order to complete tests for all students within the testing window, as well as the availability of technical staff to handle computer glitches during testing. Establishing support from key stakeholders throughout the state, including state representatives, technical staff, and test coordinators, is possible through frequent communication and a thorough training program. Such a training program should include demonstrations, simulations, and clearly defined processes for handling any alerts that occur during testing. Aside from network and system configuration, it seems likely that staff members with basic computer skills would be able to handle most of the alerts that occur during testing. Fostering partnerships between the state and testing vendor, as well as between technical staff and test coordinators, is key to increasing the comfort level of staff regarding electronic testing.

System Requirements Detail

Table 6.4

Vendors – General Population	Test Delivery Mechanism
Computerized Assessments and Learning	Smart Client
Internet Testing Systems	Browser-based
Pearson Educational Measurement Solutions	Smart Client
Vantage Learning	Browser-based

Table 6.5: Recommended System Requirements for Student Testing PCs – General Population

	Computerized Assessments and Learning	Internet Testing Systems	Pearson Educational Measurement Solutions	Vantage Learning
Processing Speed	200 MHz	233 MHz	400 MHz	1 GHz
Memory (RAM)	128 MB	128 MB	256 MB	256 MB
Available Disk Space	Not specified	Not specified	500 MB	Not specified
Display[5]	800 x 600	1024 x 768	1024 x 768	1024 x 768

[5] Display resolutions are recommended system requirements except for CAL, which is the minimum requirement.

Table 6.6: Supported Operating Systems – General Population

	Computerized Assessments and Learning	Internet Testing Systems	Pearson Educational Measurement Solutions	Vantage Learning
Windows 98	✓		✓	✓
Windows NT	✓			
Windows 2000	✓	✓	✓	✓
Windows 2003				
Windows XP	✓	✓	✓	✓
Mac OS X	✓	✓	✓	✓

Table 6.7

Vendors - Accommodated Population	Test Delivery Mechanism
Computerized Assessments and Learning	Smart Client
Measured Progress	Smart Client
Pearson Educational Measurement Solutions	Smart Client

Table 6.8: Recommended System Requirements for Student Testing PCs – Accommodated Administrations

	Computerized Assessments and Learning	Measured Progress	Pearson Educational Measurement Solutions
Processing Speed	200 MHz	1.5 GHz	1.5 GHz
Memory (RAM)	128 MB	512 MB	256 MB
Available Disk Space	Not Specified	50 MB	500 MB
Display[6]	800 x 600	1024 x 768	1024 x 768

[6] Display resolutions are recommended system requirements except for CAL, which is the minimum requirement.

Table 6.9: Supported Operating Systems – Accommodated

	Computerized Assessments and Learning	Measured Progress	Pearson Educational Measurement Solutions
Windows 98	✓		✓
Windows NT	✓		
Windows 2000	✓	✓	✓
Windows ME	✓		
Windows 2003		✓	
Windows 2000 Server			✓
Windows 2003 Server			✓
Windows XP	✓	✓	✓
Mac OS X	✓	✓	✓

Table 6.10

Vendors - Caching Server	Test Delivery Mechanism
Computerized Assessments and Learning	Smart Client
Pearson Educational Measurement Solutions	Smart Client

Table 6.11: Recommended System Requirements for Caching Servers

	Computerized Assessments and Learning	Pearson Educational Measurement Solutions – *Lab Level*	Pearson Educational Measurement Solutions – *School Level*
Processing Speed	1.2 GHz	500 MHz	1.3 GHz
Memory (RAM)	1 GB	256 MB	512 MB
Available Disk Space	512 MB (min)	Not specified	Not specified

Table 6.12: Supported Operating Systems - Caching Servers

	Computerized Assessments and Learning	Pearson Educational Measurement Solutions – *Lab Level*	Pearson Educational Measurement Solutions – *School Level*
Windows NT	✓		
Windows 2000	✓	✓	✓
Windows 2003	✓		
Windows 2000 Server		✓	✓
Windows 2003 Server		✓	✓
Windows XP	✓	✓	✓
Mac OS X	✓	✓	✓

Table 6.13: Recommended System Requirements for Northwest Evaluation Association

	MAP – Student Workstation	MAP – *Test Administration Application Workstation*	MAP – *Network Test Environment Server*
Processing Speed	100 MHz	200 MHz	2.66 MHz
Memory (RAM)	32 MB	64 MB	128 MB
Available Disk Space	Not specified	250 MB	250 MB[7]

[7] 850 MB of disk space is required when using Spanish Voice Translations.

Table 6.14: Supported Operating Systems – Northwest Evaluation Association

	MAP – Student Workstation	**MAP – *Test Administration Application Workstation***	**MAP – *Network Test Environment Server***
Windows 95	✓	✓	
Windows 98	✓	✓	
Novell 4.0			✓
Windows NT Server			✓
Windows NT	✓	✓	
Windows 2000	✓	✓	
Windows 2003			
Windows 2000 Server			✓
Windows 2003 Server		✓	✓
Windows XP	✓	✓	
Mac OS X	✓		
Mac OS X Server			✓

Best Case/Worst Case Technology Gap by Vendor

Note – all percentages are based upon computers available for testing.

Table 6.15

Computerized Assessments and Learning			
	System Requirement	**% of computers meeting requirement (best case)**	**% of computers failing requirement (worst case)**
Bandwidth	Broadband or better, T1= 96 concurrent users	56.45%	43.55%
Processing Speed	200 MHz	99.05%	0.85%
Memory (RAM)	128 MB RAM	97.46%	2.54%

In the above table, bandwidth excludes computers with dial-up and those responses of "other." Processing speed includes computers with 200 MHz and greater. Memory includes all computers with 128 MB or greater.

Table 6.16

Internet Testing Systems			
	System Requirement	**% of computers meeting requirement**	**% of computers failing requirement**
Bandwidth	Broadband or better	56.45%	43.55%
Processing Speed	233 MHz	86.19%	13.81%
Memory (RAM)	128 MB RAM	97.46%	2.54%

In the table above, bandwidth excludes computers with dial-up and those responses of "other." Processing speed includes only computers with 500 MHz and greater, as the range 200

MHz to 500 MHz doesn't allow for an estimate for computers that meet 233 MHz. Memory includes all computers with 128 MB or greater.

Table 6.17

Pearson Educational Measurement Solutions			
	System Requirement	**% of computers meeting requirement**	**% of computers failing requirement**
Bandwidth	Broadband or better, T1= 120 concurrent users	56.45%	43.55%
Processing Speed	400 MHz	86.19%	13.81%
Memory (RAM)	256 MB RAM	86.30%	13.70%

In the above table, bandwidth excludes computers with dial-up and those responses of "other." Processing speed includes only computers with 500 MHz and greater, as the range 200 MHz to 500 MHz doesn't allow for an estimate for computers that meet 400 MHz. Memory includes all computers with 256 MB or greater.

Table 6.18

Vantage Learning			
	System Requirement	**% of computers meeting requirement**	**% of computers failing requirement**
Bandwidth	Broadband or better	56.45%	43.55%
Processing Speed	1 GHz	64.24%	35.76%
Memory (RAM)	256 MB RAM	86.30%	13.70%

In the above table, bandwidth excludes computers with dial-up and those responses of "other." Processing speed includes only computers with 1 GHz or greater. Memory includes all computers with 256 MB or greater.

COMPONENT 6 – REFERENCES AND LINKS

CATS Online:

http://education.ky.gov/KDE/Administrative+Resources/Testing+and+Reporting+/District+Support/CATS+Online+Assessment/

Computerized Assessments and Learning: CAL System Requirements:

http://isat.caltesting.org/system/isat_viewall.htm

Computerized Assessments and Learning: Local Caching System Requirements:

http://isat.caltesting.org/system/lcs_system_requirements.htm

eMeasurement Services Proctor Caching User's Guide:

http://etest.pearson.com/Customers/SouthCarolina/EOCEP/resources.htm

ITS – Oklahoma School Testing Program Online Test User's Guide

One Laptop Per Child Project: http://www.laptop.org/

Measured Progress iTest System User Manual:

http://education.ky.gov/KDE/Administrative+Resources/Testing+and+Reporting+/District+Support/CATS+Online+Assessment/Online+Testing+Documents.htm

Namsoo, S., Norris, C., & Saloway, E. (2007) "Findings From Early Research on One-to-One Handheld Use in K-12 Education" (pp. 19-39) *Computing in Education: Invisible Technology, Visible Impact.* Research Center for Educational Technology, Kent State University. Mahwah, NJ: Lawrence Erlbaum Assoc.

Resolving student test alerts:

http://www.pen.k12.va.us/VDOE/Assessment/OnlineTesting/Resolving_Alerts.pdf

TESA Users Guide 2005 – 2006 (Vantage Learning):

http://www.ode.state.or.us/initiatives/tesa/tesausersguide.pdf

TestNav Technology Guidelines: http://www.pearsonaccess.com/va/support.jsp

U.S. schools may join inexpensive laptop project:

http://news.zdnet.com/2100-9584_22-6179766.html

van 't Hooft, M. and Swan, K. eds.(2007) *Ubiquitous Computing in Education: Invisible Technology, Visible Impact.* Research Center for Educational Technology, Kent State University. Mahwah, NJ.

Northwest Evaluation Association, MAP Technical Requirements:

http://www.nwea.org/assets/documentLibrary/TechRequirements1.pdf?

COMPONENTS 7 & 9 – COSTS TO THE STATE OF CONVERTING THE STATE ASSESSMENT PROGRAM TO A COMPUTER-BASED OR COMPUTER-ADAPTIVE FORMAT; COSTS TO BRING THE STATE, DISTRICTS, AND SCHOOLS TO NEEDED CAPACITY FOR DELIVERY AND MAINTENANCE OF A COMPUTER-BASED OR COMPUTER-ADAPTIVE STATE ASSESSMENT

Cost is a key driver in the decision whether to move a state assessment program from paper/pencil to a computer-based or computer-adaptive delivery. States operate under budgetary constraints and must prioritize where the limited funds will go. Determining an overall cost to convert the complete assessment program for the state of South Carolina to electronic delivery is dependent on many variables. For example, the selection of a vendor or vendors, the aggressiveness of the implementation plan, and changes to the testing design itself all contribute to the development of an overall budget. Thus, this section will provide budgetary estimates and a detailed overview of the costed components. This begins with a discussion and an explanation of why electronic testing is not initially a cost savings over paper/pencil testing nor does it necessarily ensure savings over time.

The Myth of Immediate Cost Savings

Many states approach the conversion of their assessment programs from paper/pencil testing to electronic delivery with the notion that there will be dramatic and immediate cost savings to the program. This belief hinges on the idea that with a reduced or eliminated need to print, package, ship, and scan the volumes of paper currently required that costs must surely decrease. However, experiences in states that have successfully converted their assessment delivery methods to electronic delivery have not shown this to be true. In fact, all states interviewed as a part of this study have shown similar cost patterns in their programs. What has emerged is that, in the initial one to three years of conversion, states see a significant upward spike in the total cost for test delivery. This spike can be attributed to a number of factors.

- **Costs of infrastructure improvement** — The conversion to an electronic delivery assessment model almost always means that a significant, up-front investment in technology infrastructure (computers, servers, bandwidth, etc.) must be made in testing locations to support the number of computers necessary to achieve the goals. The larger the scope of the testing program and the more aggressive the implementation schedule, the larger the up-front investment.

- **Dual implementation costs** — Rarely do states move their entire testing program from paper/pencil to electronic in a one- or two-year time period. More commonly this is accomplished over a three- to six-year period. This means that for a significant time period, the program must support the processes and associated costs of both paper/pencil and electronic delivery with an increase in total cost. There are a number of logical reason for this taking this approach so the dual mode may continue for a lengthy period.

- **Fixed-cost nature of electronic testing** — In an electronic delivery testing model, a higher percentage of the total vendor cost is fixed, meaning that the cost to develop software to accommodate electronic student registration, test delivery, scoring, and reporting is the same regardless of the number of students tested. Most implementation plans recommend a relatively small student participation rate in the early years to allow infrastructure, training, and delivery processes time to mature. While these are all valid reasons to proceed slowly with implementation, the result is a higher cost per student than it would have been if participation rates were initially higher.

- **Cost of comparability studies** — Because most program conversions move slowly in the early years, individual grades and content areas have a blend of both paper/pencil and electronic test delivery. In order to meet NCLB regulations, a dual mode approach requires additional psychometric analysis in the form of comparability studies to ensure equality across the modes. While psychometric adjustments can be made to make the modes comparable under a post-equating model, the dual analyses would need to be conducted each year rather than once, as would be the case for a complete conversion. These are costs that would not be necessary in a single-mode delivery model.

Our review of other states that have gone through this process shows the initial upward spike in cost peaks and eventually decreases as more of the testing program is converted to an electronic delivery model. At the Expert Panel discussion conducted for this study, representatives from the state of Idaho indicated that it took them four to five years to begin achieving the kind of return on investment that they had originally expected when converting their assessment program. This experience was echoed by representatives from other states as well. It should also be noted that every state participating in the Expert Panel indicated that despite the initial cost, they felt strongly that converting their programs to electronic delivery was a sound investment in the future of the programs and the overall education of students.

Cost Components in Converting from Paper Testing to Electronic Testing

This section provides additional detail regarding the cost components involved in converting to an electronic testing system. The purpose of this discussion is to provide insights into the incremental cost differences (increases or decreases) of electronic testing versus paper testing. The assumption is that there are currently paper-testing assessments in place, and the analysis considers the changes required to convert to electronic testing. This study explains the following cost categories or drivers and how they impact total cost of a testing program when moving from paper/pencil delivery to electronic delivery.

Table 7.1: Cost Categories and Their Effect on Vendor and State Costs

Category or Driver	Vendor Cost	State/District/ School Cost
Computers, hardware, and bandwidth	↑ Software development	↑ New computers Computer upgrades Increased internet bandwidth
Item development	↑ Reformatting existing items for electronic delivery Expansion of item bank Innovative item types	
Training	↑ Additional training materials Additional training delivery	↑ Additional training of technology coordinators

Table 7.1 (continued): Cost Categories and Their Effect on Vendor and State Costs

Category or Driver	Vendor Cost	State/District/ School Cost
Staffing		↑ Involvement of technology coordinators Increased proctoring
Adjustments to computer labs		↑ Reconfiguration of seating Partitions or privacy filters
Printing, packaging, and shipping of test materials	↓ Reduction in printed test materials	↓ Reduced time spent handling printed test materials
Psychometric analysis	↑ Comparability studies	↑ Sampling and administrative

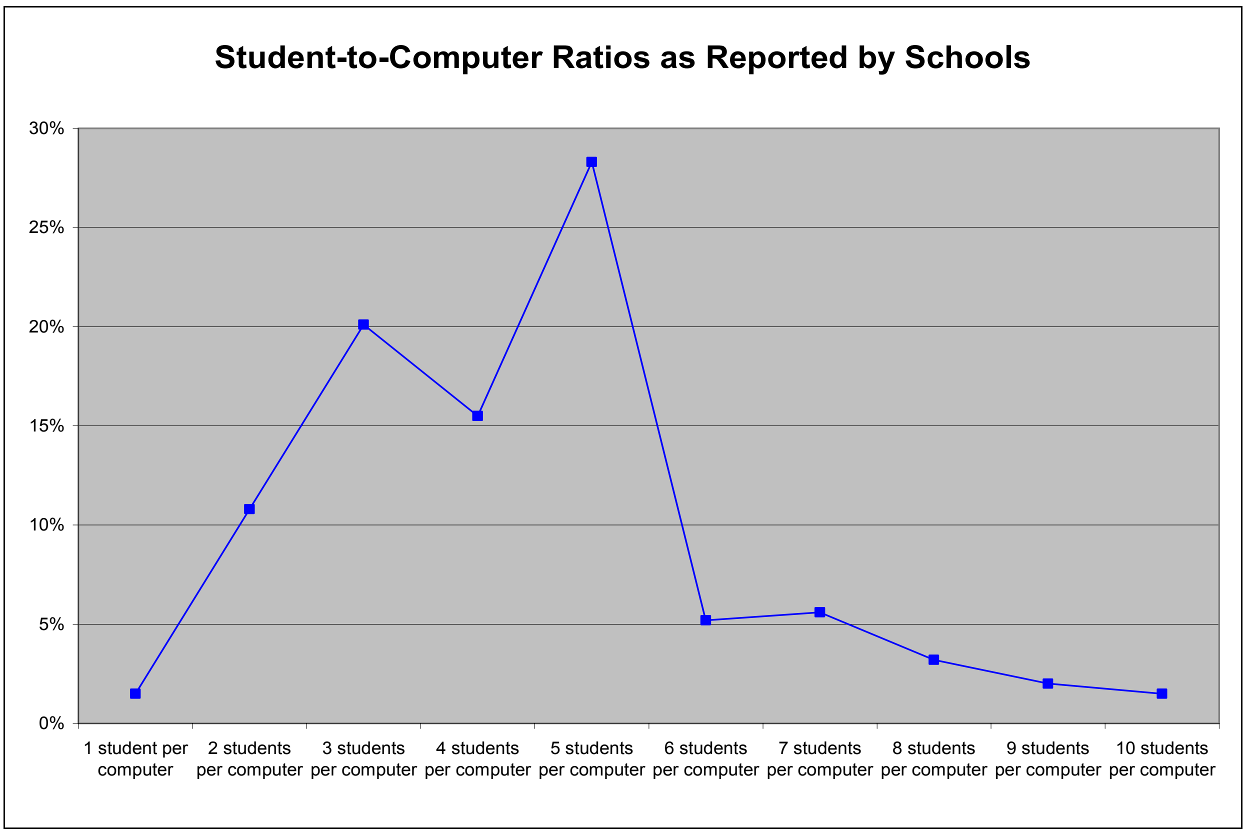
Cost Categories/Drivers

HARDWARE, SOFTWARE, AND BANDWIDTH IMPROVEMENTS

This represents the cost necessary to close any technology gap that could negatively impact the successful implementation of an electronic testing program. The cost required to close the technology gap becomes a function of how aggressively the state would choose to implement electronic testing. The most commonly cited factor to accomplish this is the ratio of students to computers. In states such as Idaho, Oregon, and Virginia, which currently conduct a majority of their large-scale assessment programs via electronic testing, the student-to-computer ratios are between 3:1 and 3.5:1.

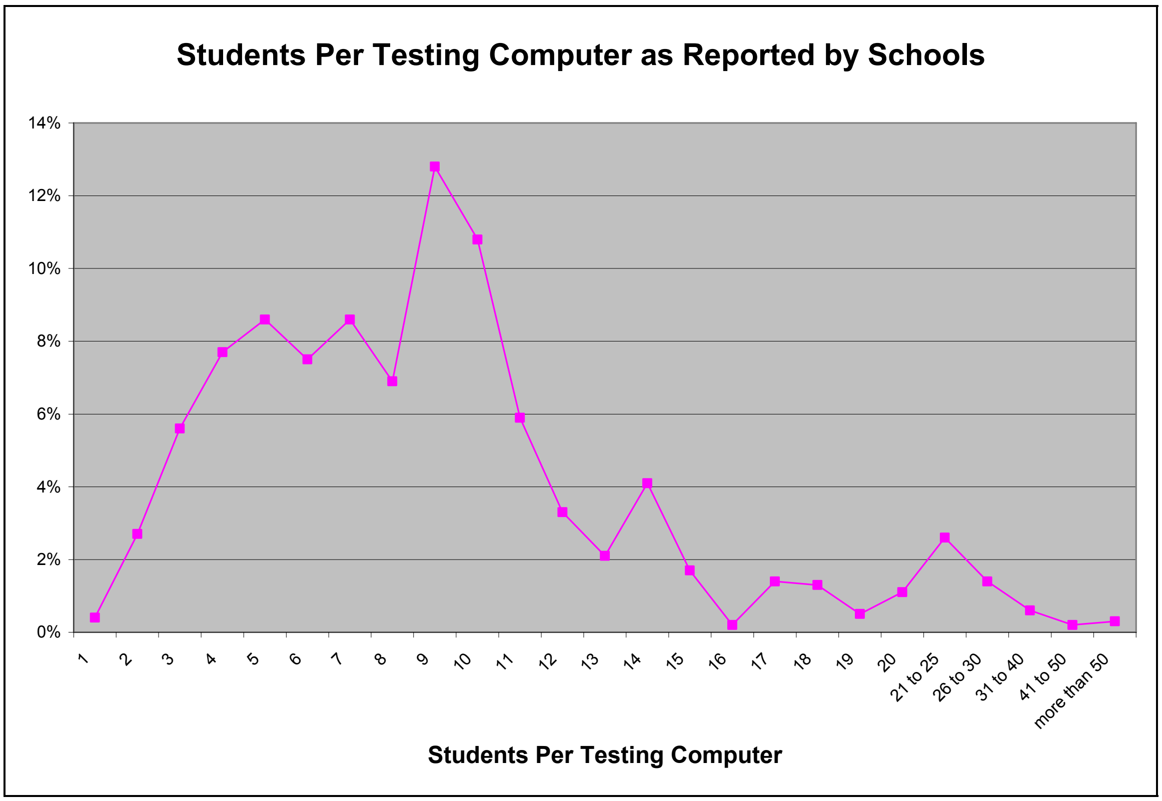
This study's survey of South Carolina educators yielded two significant pieces of information about student-to-computer ratios. The survey responses indicate there are a total of 177,604 computers available to students in schools across the state. Given an estimated student population of 650,000 (50,000 students per grade level) this yields an average student-to-computer ratio of 3.7:1 across the state. Note that the ratio for a particular school will differ. The most common response by South Carolina educators was a student-to-computer ratio of 5:1.

Figure 7.1



However, the results become more disparate when evaluating them against the number of computers that could be used for testing purposes instead of the total of all computers available to students. As discussed in Component 6 of the study, the survey responses indicated that 100,372 computers, or 56% of the total computers available to students (see Figure 6.3), could be made available for assessment purposes. For the remainder of this component, we will refer to this number as reported by South Carolina schools as the students-to-testing computer ratio.

Figure 7.2



In this case, we see that 16.3% of schools indicated that they have a student-to-testing computer ratio better than 4:1. These results indicate that the vast majority of schools would need to acquire a significant number of additional computers and related hardware to successfully implement electronic testing on a large scale. This number is significantly higher than student-to-computer ratios in most, if not all, states currently delivering a majority of their high-stakes summative testing programs online. Without an improvement in the student-to-testing computer ratio, it would not be feasible for most schools to complete testing within a reasonable testing window. This study recommends that the state work towards achieving a minimum student-to-testing computer ratio of 4:1. Using the estimated student population of 650,000 as discussed above, South Carolina would need a total of 162,500 computers available for testing purposes, or 62,128 more than are currently available.

Table 7.2: Total Computers Needed to Accommodate Electronic Testing

Total computers to achieve a 4:1 student-to-testing computer ratio	162,500
Total computer available for testing (based on survey data)	100,372
Number of additional computers needed	**62,128**

The 62,128 computers needed represents an overall number and could be achieved either by purchasing these computers in the open market or by reallocating a percentage of the computers currently in the schools that are not available for testing and making them available for testing or some combination of the these two approaches. While it is unlikely that the entire difference in the number of computers could be made up by reallocating currently owned hardware, this study estimates that it would be both possible and feasible to reallocate a significant percentage of these machines during a limited testing window so as to minimize the total infrastructure investment necessary. This study would recommend meeting this additional computer need as outlined in Table 7.3.

Table 7.3: Recommended Method for Meeting Additional Computer Need

Purchase/acquisition of new computers	32,500
Reallocation of currently owned computers	29,628
Number of additional computers needed	**62,128**

To begin to determine an overall cost for infrastructure investment, this study first takes the recommended 32,500 additional computers to be purchased and applies an estimated purchase and installation cost of $1,000 per workstation. This estimated per unit cost is based on current market research. This would mean an investment of $32,500,000 in computer acquisition costs.

The acquisition of 32,500 addition computers for schools would necessitate investment in other supporting devices as well. To make certain that these computers where available for use in both assessment and instruction, other infrastructure upgrades including supporting servers, cabling, power, and bandwidth would be necessary. For purposes of this study, we estimate these supporting costs to add 30%, making total cost for the acquisition of new hardware and infrastructure to be approximately $42,250,000.

This number represents the cost of new hardware and bandwidth that would be required to meet the target student-to-testing computer ratio. It is assumed that all new hardware purchased would meet the recommended system requirements as defined by the assessment vendor. However, the survey results and analysis described in Component 6 also show that a significant number of existing computers would need upgrades to meet the base vendor requirements. Because the recommended system requirements can vary significantly from one vendor to the next, it is difficult to determine an exact number of upgrades that would be necessary. This study recommends that the state work closely with its chosen testing vendor to determine any system upgrades that would be necessary. For budgetary purposes, we will assume that 30% of the 100,372 computers currently available for testing in the state will require some form of upgrade (memory, processor, operating system, etc.) at an average cost of $400 per computer. This would result in upgrade costs of approximately $12,000,000. This would bring the total hardware and infrastructure investment to an estimated $54,250,000. A summary of the recommended infrastructure investment is presented in Table 7.4.

Table 7.4: Summary of Infrastructure Cost

Cost Item	Amount
Acquisition of additional computers	$32,500,000
Acquisition of additional hardware (servers, networking equipment, cabling, power)	$9,750,000
Upgrades to existing computers (memory, processor, operating system) to meet recommended system requirements for test delivery	$12,000,000
Total recommended infrastructure investment	**$54,250,000**

This recommended investment is intended to represent the total cost for hardware and infrastructure needed to bring an entire assessment program online. It is not intended to imply that all of this amount must be invested before any part of electronic testing could begin. This cost would be spread over time depending on the aggressiveness of the defined implementation plan. The amount of necessary investment would follow in close proportion to the percent of students participating in the online program each successive year.

Beyond the initial investment, there will be ongoing annual costs to maintain, repair, and replace the hardware and infrastructure. Technology advances rapidly, and obsolescence in the personal computer market can appear to happen overnight. This study was unable to determine whether any statewide policy for a computer replacement schedule exists in the state of South Carolina. In order to keep up with rapidly changing environments and technological advances, this study recommends that the average useful life of a computer to be used for electronic test delivery is no more than five years. This implies that 20% of the average school computer fleet should be replaced on an annual basis. *These costs are not factored into the estimates provided in Table 7.4*. However, they should be considered as a real and direct cost of this type of investment. Funding for these ongoing costs presents an understandable challenge for a majority of districts and schools.

It should be noted that the number of computers and student-to-computer ratios derived from the survey of South Carolina testing locations for this study differs significantly from what has been reported in *Education Week*'s "Technology Counts 2007: A Digital Decade" surveys. This study has not attempted to determine the causes for these discrepancies. However, we believe that the specific ratio of student-to-testing computer used here is a more realistic definition for the desired student-to-computer ratio. That is, it represents the survey respondent's estimate of computers available specifically for testing and therefore presents a lower boundary or worst-case estimate. That is, if South Carolina achieved the same ratios as reported by other states conducting high stakes testing, it is highly likely that the resources would be sufficient.

The cost estimates herein provided for technology infrastructure are conservative and should be considered budgetary guidelines only. The cost of technology changes rapidly. Increases or decreases in hardware purchase and upgrade costs will likely have a significant impact on the cost figures presented. Additionally, states may be able to achieve significant discounts in the per-unit cost of technology hardware through educational discounts and/or bulk purchasing. These discounts would also impact the total dollar investment needed for hardware and infrastructure.

ITEM DEVELOPMENT

When moving from a paper/pencil assessment to a computer-based or computer-adaptive model, there are a number of costs that may be incurred with the items associated with the program. Examples of these include:

- Item conversion/formatting — Existing items may need to be reformatted in order to make them appear as desired on the computer.
- Expansion of item bank — Moving from paper/pencil to computer-delivered tests often means expanding the testing window to accommodate the number of students taking the tests. A common method for addressing the added item security concerns of an expanded testing window is through the use of multiple test forms. However, multiple test forms result in more testing items and hence additional item development costs. Also, programs that wish to use a computer-adaptive testing model will typically require significantly larger item pools. This also results in higher item development costs.
- New item types — A potential benefit of electronic testing is the increased potential for innovative item types such as simulations and streaming audio/video. The licensure and certification markets have successfully implemented many new item types. These new item types have the potential to allow testing to occur in ways not possible on paper. However, these item types come with a significantly higher item development cost.

This study has made no assumptions as to what items the state of South Carolina would include in computerized testing (e.g., existing items, newly developed items, items adopted from a vendor). Therefore, item or test development costs have not been considered.

TRAINING

Moving to an electronic testing environment would require new and additional training of staff. Training of technology coordinators and test administrators on the delivery engine is often a cost that is not incurred in the paper/pencil environment. For example, technology coordinators would need to be trained on the delivery engine and its associated system/server requirements, and the installation and updating process. Test administrators would also need training on the test engine. This would be in addition to any ongoing training for paper-pencil testing should South Carolina have concurrent paper/pencil and computer-based or computer-adaptive testing programs during a transition period. In most cases, this is not seen as a direct cost to districts or schools as the training is provided by the testing vendor and included in the overall cost of the assessment program to the state. However, there may be an opportunity cost in having technology coordinators away from the office to attend training for their new role in electronic statewide testing.

When surveyed, South Carolina educators estimated they currently train 3.38 test administrators and 1.02 technology coordinators per school as a part of the assessment programs. These same educators felt that if the program were moved to an electronic test delivery system, they would need to train 3.51 test administrators and 2.46 technology coordinators per school.

STAFFING

The movement from paper/pencil to electronic testing often requires modifications to resource staffing during the testing window. As students are rotated in and out of computer labs for testing, it may require additional staff to monitor the students in the lab and in the classroom at the same time. Also, districts and schools may need greater access to technical support staff during the testing window. A detailed analysis of the staffing and training implications is covered in Component 8 of this study.

ADJUSTMENTS TO COMPUTER LABS

Most electronic assessments in schools today take place in computer labs. However, this is not an ideal environment as these labs are rarely arranged with high-stakes assessment in mind. For example, computers are often arranged close together with limited desk space. Also, the upright nature of computer monitors allows a student to see the screen and responses of a neighbor's computer more easily than in a paper/pencil environment. The purchase and installation of privacy filters on monitors are a means of addressing this increased potential for student cheating, but this can be a costly proposition. These filters currently range in price from $100 to $250 per unit. That means a school with 50 computers to be used for testing would need to invest between $5,000 and $12,500 to mitigate this risk. Component 10 offers recommendations that have been implemented in other states to reduce this risk at a lower cost.

Another factor to consider is that many computer labs do not have phones installed at all or the existing phones are "blocked" to prevent calls to external phone numbers. In the event of a system issue, it will be necessary for the test administrator in the lab to contact technical support over the phone. More often than not, the administrator will need to be in front of the computer to describe the problem, thus making phone access in computer labs a needed requirement.

The implementation of wireless networking technologies and the increased use of mobile computer labs or computer carts may provide the best solution to the configuration issues noted above. These carts typically contain 24 to 36 laptop computers that can be brought into a classroom of students for instructional or assessment purposes. The computers use secure wireless networking to connect to the Internet. Bringing the computers to the student classrooms has two major benefits:

- Standard classroom configurations with rows of desks provide a better arrangement for testing.
- Having students take an electronic test in the same classroom as they would a paper/pencil test lessens the impact on regularly scheduled instructional time in the computer lab.

For these reasons, the state of South Carolina should strongly consider a mobile-cart model for a significant portion of the new computers that will be needed for an electronic testing initiative. Laptops and mobile carts provide a great deal of flexibility to students and teachers for both assessment and instructional purposes. Laptops, however, are generally more expensive to purchase than desktops, and typically have a shorter life.

Printing, Packaging, and Shipping of Test Materials

By far the most significant cost savings to any state moving an assessment program from paper/pencil to electronic delivery comes in the reduction or elimination of paper testing materials. Paper testing materials must be printed, packaged, distributed, received, and scanned by the testing vendor(s). This constant handling of high volumes of paper represents a large number of labor hours by testing vendors. Additionally, districts and schools must have personnel tasked to manage the paper materials during the administration window. This often includes verifying shipped quantities, distribution to classrooms, collection, and repackaging for return. With the conversion to electronic delivery of the tests, the total amount of paper assessment products can be drastically reduced but not eliminated.

PSYCHOMETRIC ANALYSIS

Any significant change to a large-scale assessment program should be carefully analyzed to determine its potential impact on student results. The conversion from a paper/pencil to an electronic delivery testing system is no different. The early stages of implementing any conversion plan should include detailed comparability studies and analysis of testing data by trained psychometricians. Examples of these types of studies can be found in Component 4 of this study. Other studies investigating test anxiety and computer experience (or lack thereof) should also be conducted. The cost of this additional work typically increases the total cost of the program in the early years of implementation.

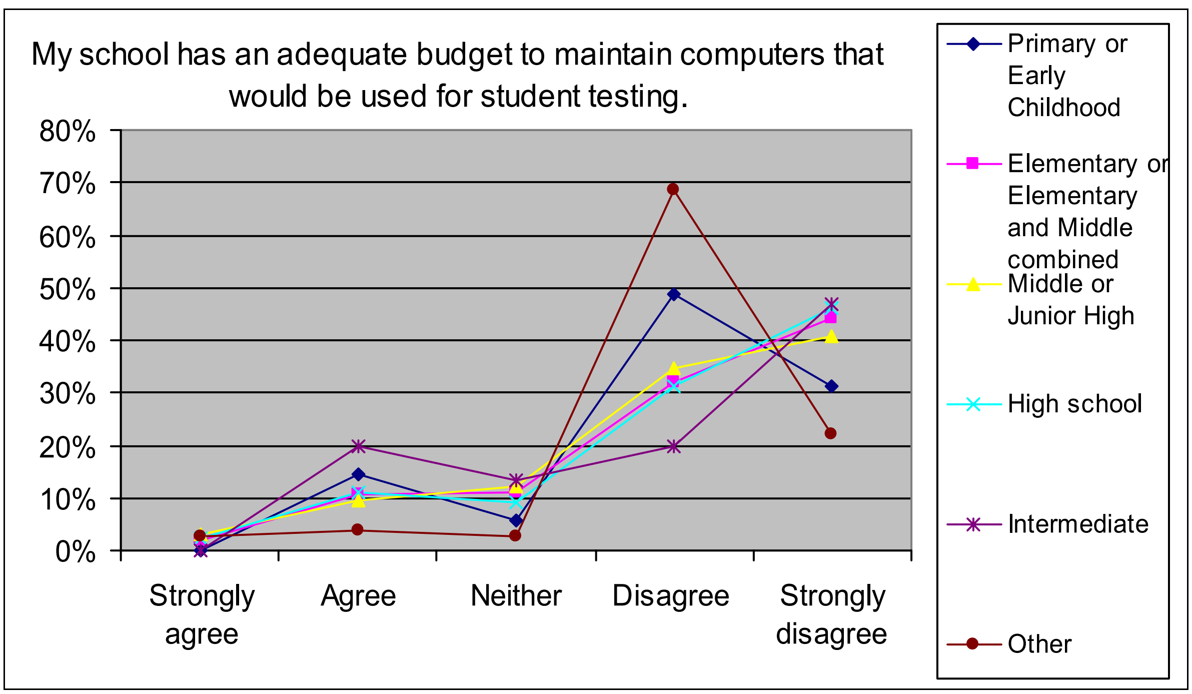
Potential Costs to the State of South Carolina

The potential costs to South Carolina depend on the implementation model or path that is chosen. That model will determine the cost components described in this section. For example, once the plan is defined for the number of online testers each year, the computer costs, computer upgrades, bandwidth costs, staffing, and training costs at the schools and districts can be plotted against the implementation timeline. The estimates for these costs for South Carolina can best be shown as part of the recommended implementation plan (see Components 13 & 14).

Funding Perceptions Across the State

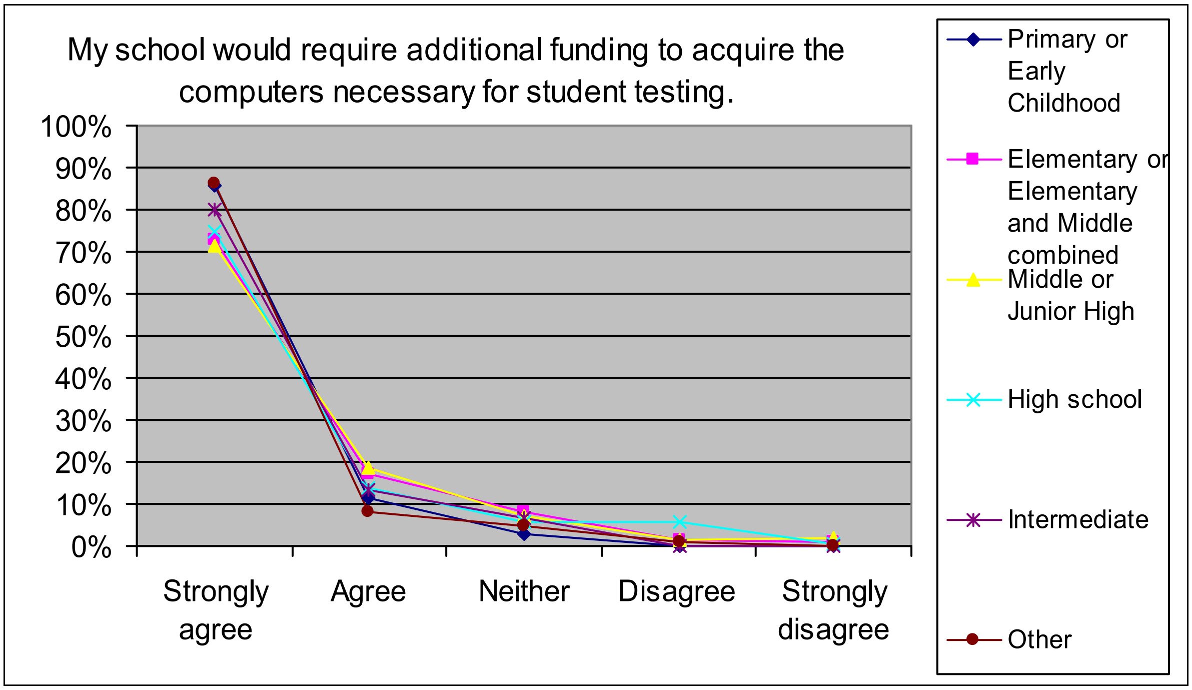
If the state of South Carolina moves toward additional electronic testing, one immediate hurdle to be overcome will be the perception in the field about the school-level budgets necessary to accommodate electronic assessment. Currently there is a predominant perception in the field that electronic testing will mean increased costs for the schools. Within the survey, South Carolina educators were asked whether they felt their schools had adequate budgets to maintain the computers that would be necessary for electronic testing. Seventy-eight (78%) percent of respondents indicated that they either "disagree" or "strongly disagree" with the statement, "My school has an adequate budget to maintain computers that would be used for testing." The following figure shows the findings broken down by primary, elementary, junior/middle, high school, intermediate, and other schools:

Figure 7.3



The study also asked respondents to indicate whether they felt additional funding would be necessary to acquire computers to be used for electronic testing. An overwhelming majority (90%) responded that they would require additional funding to move to electronic testing. The same school type breakdown as Figure 7.3 is provided below.

Figure 7.4



These results show a strong agreement among test administrators that their school would require additional funding to acquire adequate computers for electronic testing.

Summary

When any state is considering a transition to electronically delivered assessments, cost always becomes a key factor. The desire to implement technology advancements to provide the best and most useful testing environment for students must be balanced with budgetary considerations. A common misperception is that the reduction in printing, packaging, and distribution costs associated with a paper-based assessment program will lead to immediate and dramatic cost reductions. However, experience has shown this not to be the case because of the introduction of new costs such as infrastructure upgrades, increased item development costs, and additional comparability studies to meet NCLB guidelines. All of the states surveyed as part of this study showed an initial spike in costs during the first years of implementation.

The total cost of implementation for any state depends on a multitude of factors that make it virtually impossible to estimate total cost precisely. This presents both a challenge and an opportunity for budget-driven stakeholders. While there is a challenge in attempting to develop a budgetary number for appropriation purposes, the state can have a large amount of flexibility in tailoring the aggressiveness of an implementation plan to the funding available. More aggressive plans will come with both higher up-front costs and faster progress toward cost savings. A less aggressive approach will allow the technology investment to be spread out over a longer period of time but will likely entail a longer return-on-investment period.

In the state of South Carolina, cost is clearly a major concern for staff in the districts and schools. Virtually every testing site felt they would need additional funding for technology and staff to be able to accommodate a transition from the current delivery model to an electronic one. To counteract this perception, the state will need to establish clear policies and communication plans with test coordinators and administrators to fully understand and address the needs of individual sites. The state will need to assist districts that have been slower to keep up with modern technology.

COMPONENT 7 & 9 – REFERENCES AND LINKS

"Technology Counts 2007: A Digital Decade," *Education Week* (Vol. 26, Issue 30) March 29, 2007. Available at http://www.edweek.org/ew/toc/2007/03/29/index.html

COMPONENT 8 – CURRENT STATE, DISTRICT, AND SCHOOL CAPACITY, TO INCLUDE PERSONNEL, FOR ADMINISTERING A COMPUTER-BASED OR COMPUTER-ADAPTIVE STATE ASSESSMENT

A detailed analysis focusing specifically on the personnel needed to implement and administer computerized testing is profiled below. The profile was created using the Adjusted[1] Feasibility Study Survey dataset.

Capacity Profile – District

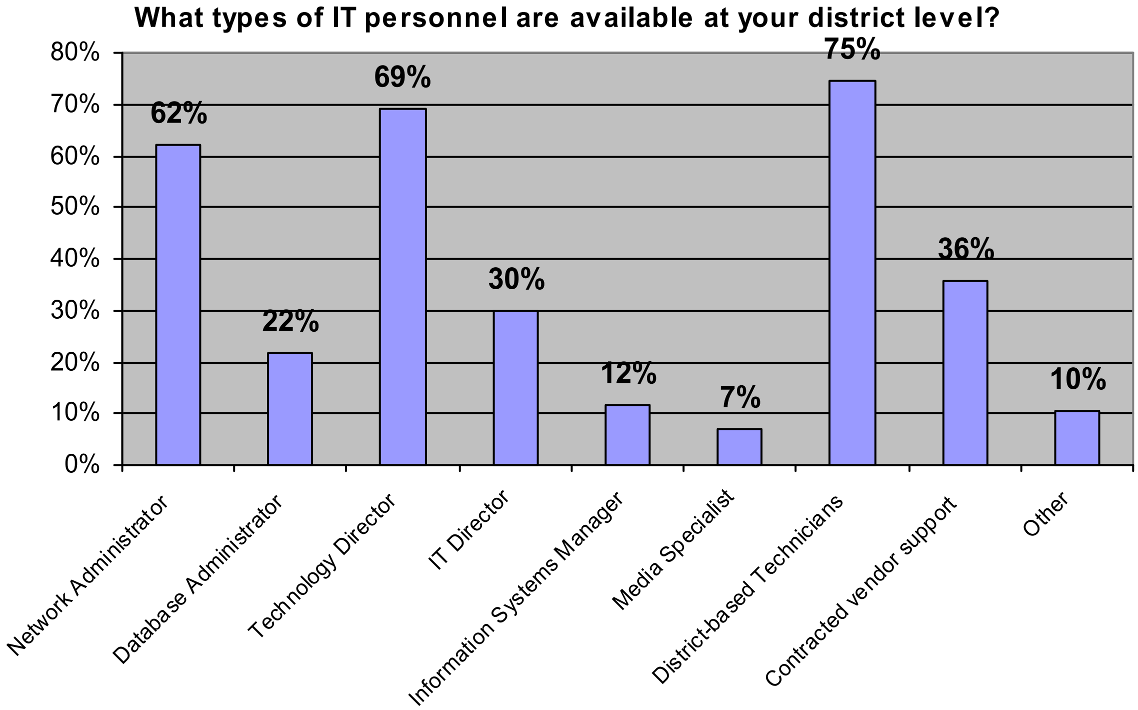
The survey included a set of questions specifically aimed at profiling the staffing/training resources at the district level. Following are highlights of the findings.

Implementation of computerized testing would involve addressing a number of issues at the district level (e.g., installing software, ensuring district systems are compatible with the testing software, ensuring sufficient bandwidth exists). Survey results indicate that most districts have at least one information technology (IT) resource to draw upon for these tasks.

The average number of IT staff available at the district level is five. As displayed in Figure 8.1, the majority of South Carolina school districts have network administrators, technology directors, or district-based technicians available to them.

[1] Adjusted dataset - A complete set of data including every district and school was not obtained because 71% of the districts and schools responded to the survey. The robustness of the obtained data allowed survey responses to be estimated for the non-responding districts and schools. For more details on the methodology used for creating the adjusted dataset, please see the Adjusted Dataset portion of Survey Methodology & Overview section of this report.

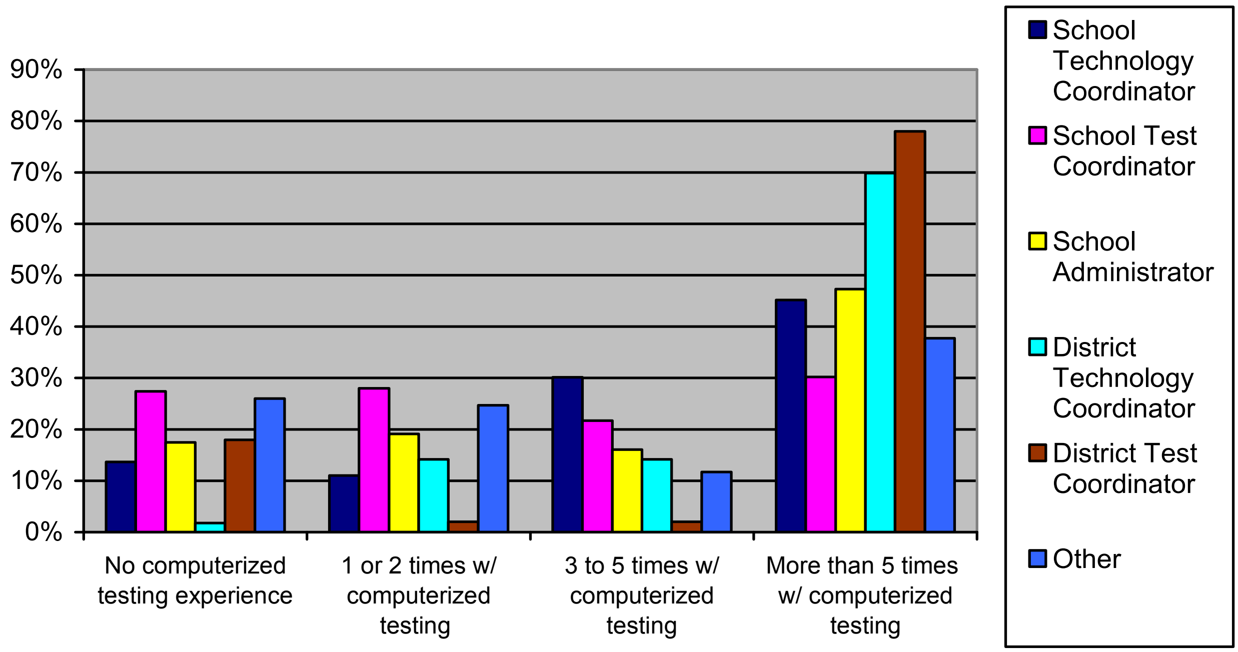
Figure 8.1: Types of IT Personnel Available at the District Level



While there is a wide range of experience among technical and testing staff (see Figure 8.2), the majority of these individuals have had some level of experience with computerized testing (one or more instances). District test coordinators and district technology coordinators appear to have the most experience with computerized testing.

A significant percentage of these individuals at the district level have had extensive experience (i.e., 5 or more times), suggesting that South Carolina has a pool of experienced district individuals to draw from during an implementation phase. If South Carolina moves forward in computerized testing, districts with a higher number of IT personnel and schools where technology is an integral part of daily instruction should be studied to identify best practices. Alternatively, districts with personnel who have lower levels of experience might be paired with districts with higher levels of experience in a peer mentor or resource program. Finally, depending on how South Carolina chooses to proceed, those districts with more extensive experience should be tapped for task forces that would help with recommendations and planning for implementation of computerized testing.

Figure 8.2: Level of Experience with Computerized Testing



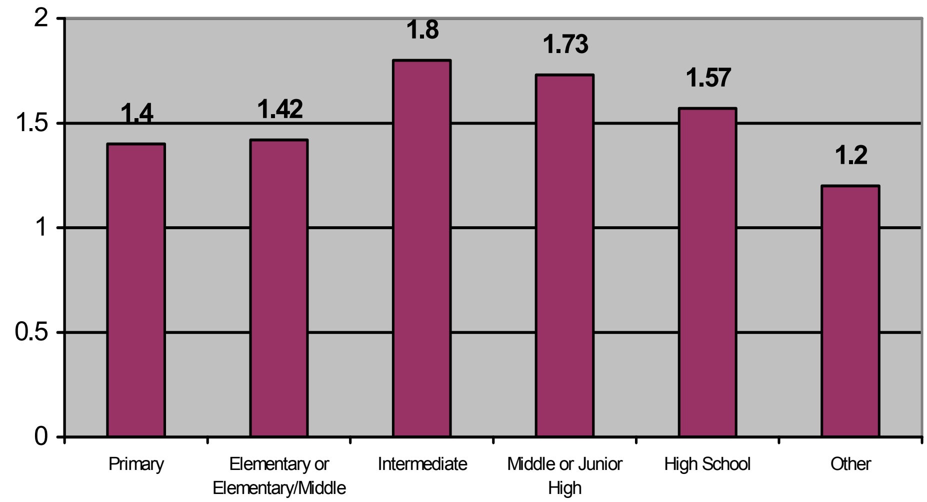
While the majority of district technology coordinators and district test coordinators indicated on the survey that they had worked with computerized testing "more than 5 times," school-level personnel have less extensive experience. Fifty percent (50%) or more of the school technology coordinators, school test coordinators, and school administrators have three or more experiences with computerized testing. Finally, it is notable that the percentage of districts indicating that technical and testing personnel have had no experience whatsoever is relatively low. This implies that training for most district technical and testing personnel would be manageable.

Capacity Profile – School

The survey included a set of questions specifically aimed at profiling the staffing/training resources at the school level.

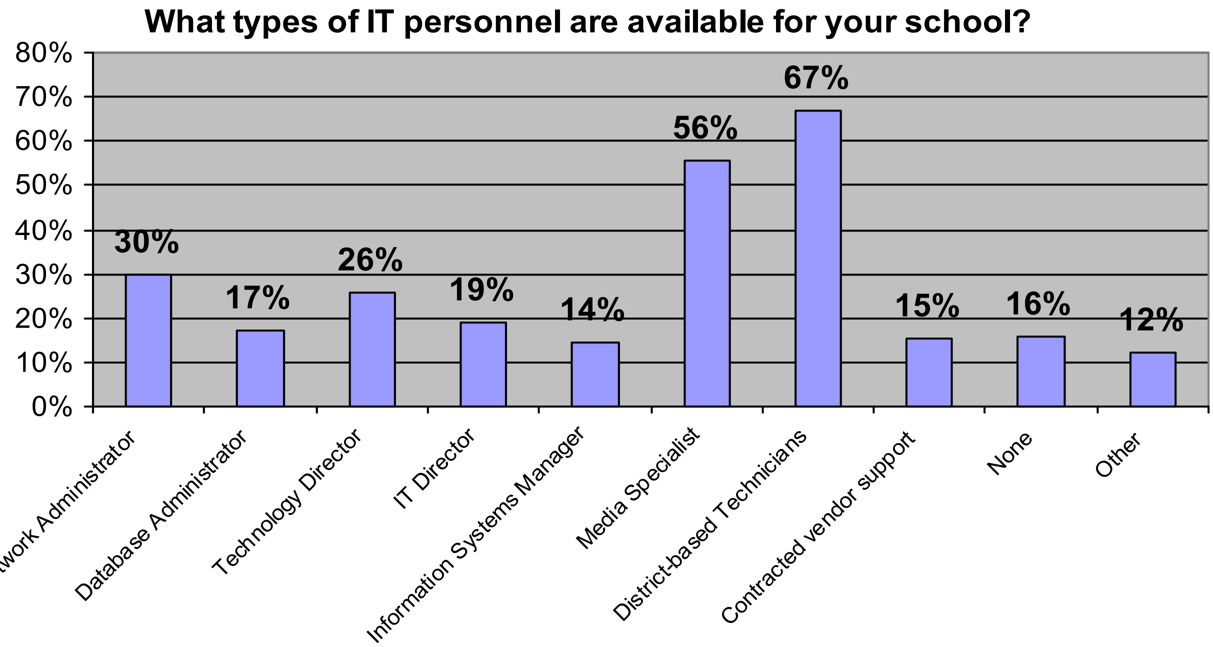
The survey findings showed that the average number of IT staff available at the school level was 1.5. The statewide average number of IT staff varies across type of school, peaking with intermediate or middle/junior high.

Figure 8.3: Average Number of IT Staff Available for Each School Type



School-level IT personnel are predominantly media specialists (56%) and district-based technicians (67%). This finding is consistent across all types of schools. The finding that district-based technicians are available to most schools is positive, as these resources should be called upon to help with issues of initial implementation (e.g., software installation). However, it is also likely that district-based technicians provide service to several different schools. Consequently, South Carolina would need to take into account the number of schools technicians are expected to support when determining if and how to implement computerized testing.

Figure 8.4: Types of IT Personnel Available at School

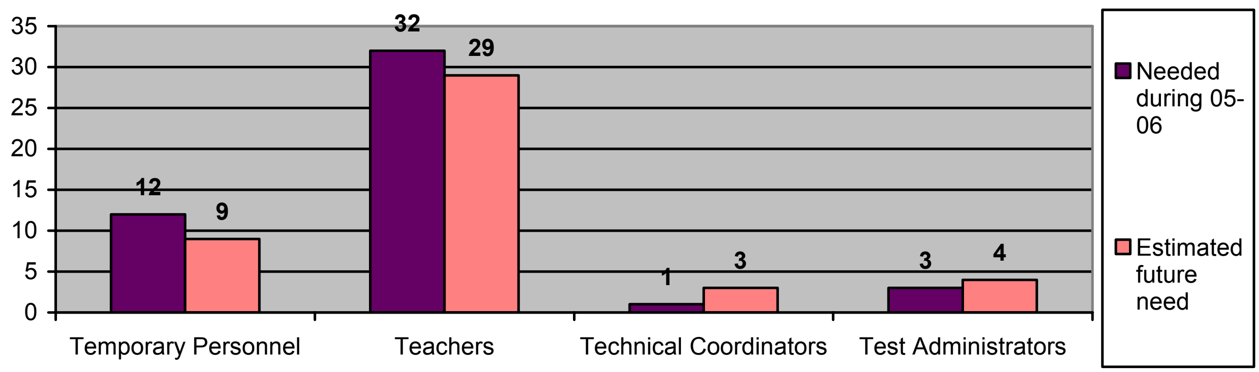


Survey questions also addressed the need for temporary personnel to provide assistance during the test administration. Survey results showed that the average number of temporary personnel needed during the 2005–2006 test administrations was 12 per school. The average number of teachers trained annually for test administration was 32 per school; the average number of technical coordinators trained was one per school; and the number of test administrators was three per school.

Survey respondents were also asked to provide their estimates of personnel needs if the school were conducting computerized testing. (Note: only those survey respondents having computerized testing experience (83%) provided their estimates.) The average number of temporary personnel needed for computerized test administration was estimated to be nine per school, and the number of teachers needed was estimated at 29 per school. The estimated number of technical coordinators was three per school and the number of test administrators is four per school.

A comparison of these two sets of personnel estimates (2005–2006 and if computerized testing were utilized) shows that survey respondents surmised that computerized testing would require somewhat fewer temporary personnel and teachers to be trained under computerized testing conditions. However, the respondents estimated that more technical coordinators and test administrators would be required at the school level for computerized testing.

Figure 8.5: Comparison of Personnel Needed if Utilizing Computer Testing



Resources Required to Address Complexity of Implementation

Implementation of online tests can be complex because both the technology and the assessment elements of the testing process have to be aligned and working well.

Personnel in charge of supporting the online testing environment will need to have the appropriate skills to address the following challenges:

- Addressing hardware and software requirements delineated in Component 6 (memory RAM, processing speed, bandwidth, operating systems);
- Preparing local infrastructure for testing readiness (system requirements, needed computer upgrades, and sufficient testing of local software so that it will not interfere with the testing system);

- Facilitating the administration of the online test and monitoring the testing lab or classroom environment to ensure computer readiness, test security, and test administration issues. For example, test-day issues may include system re-boots, error messages, or other unforeseen issues); and

- Monitoring the online test administration to ensure that appropriate testing procedures are followed and that test security is maintained.

Valuable insights about personnel needs and issues were offered by representatives of eight states who participated in the SC Feasibility Study Expert Panel on March 28, 2007, in Columbia. The following are key points regarding personnel, technical support, and lessons learned:

- Training for staff and students is critical. State representatives must participate in state training sessions to answer policy questions. (Idaho)

- There are two elements of training – adult training (administrative) and student training (how do I navigate through the test and use the testing tools). (Virginia)

- It is not the hardware that is the problem, but having people with expertise and technical skills to implement computerized testing. (Oklahoma)

- Technology, testing, and education experience and expertise are all needed. All teams should have these three audiences at the table. (J. Poggio)

- Instruction will drive success in computerized testing. Teachers must be comfortable using computers in instruction. Teachers should be provided professional development to do so. (West Virginia)

- Make sure instruction matches testing. (Virginia)

The above points highlight resource training, skill sets, knowledge and comfort level with online testing, and addressing challenges with the right groups of people.

Link to Instruction

An important element in successful computerized testing is the alignment of assessment and instruction. The assessment venue should reflect the mode of delivery of instruction. As noted above, advice from the Expert Panel affirmed that providing instruction solely in one medium while assessing in another should be minimized or avoided so that student performance on the assessment accurately reflects student learning. An example of linking assessment to instruction can be illustrated in mathematics. If students are permitted to use calculators during instruction, they should be permitted to use online calculators during the assessment. Likewise, if in science students create and manipulate graphs to display their findings, they should have a similar experience during the assessment.

Assessment, Technology & Education Partnership

Critical to the success of administering online assessments is the partnership among the assessment, technology, and education representatives within the state. Each group plays an important role in developing and implementing successful, user-friendly, and instructionally relevant online testing. Each group brings a different perspective to the table, all of which are critical to the shaping of a successful computerized system. Both Virginia and Kansas (J. Poggio) emphasized the importance of this strong link between the three groups at the Expert Panel meeting.

SUMMARY

South Carolina is developing a statewide infrastructure of state, district, and school personnel who have expertise in implementing and maintaining robust technology systems that can support both computerized testing and well-grounded computer-enhanced instruction. This infrastructure should continue to grow incrementally in numbers and mature in expertise as a computerized testing program is implemented in South Carolina. In this manner, the state can support the integration of assessment and instruction so that the resources allocated to computerized testing can also serve to support computer-enhanced instruction.

The specifics of any operational planning for personnel to support a conversion to computerized testing depends in large part on the specific factors and requirements of the plan chosen (e.g., specific testing system chosen, aggressiveness of the implementation plan). However, the following summary points are worth noting as implementation plans are considered:

- There is a wide range of experience levels among technical and testing staff, but the vast majority had some level of experience with computerized testing (one or more instances). District test coordinators and district technology coordinators have the most experience.
- Survey responses indicate that South Carolina has a pool of district-level individuals with computerized testing experienced to draw from for future action planning, training, etc. (i.e., 70% or more of district technology coordinators or district test coordinators have had five or more experiences with computerized testing). The level of experience is lower at the school level, but 50% or more of school technology

coordinators, school test coordinators, and school administrators have had three or more computerized testing experiences.

- For computerized testing, personnel needs for testing setup and administration are likely to shift. Responses indicate that computerized testing would require fewer temporary personnel, and fewer teachers would need to be trained in testing. However, respondents estimated that more technical coordinators and test administrators would be required at the school level.
- Responses show that district-based technicians are available to most schools. However, it is likely that district-based technicians would provide service to several different schools. Therefore, if a decision is made to implement computerized testing, South Carolina would need to take into account the number of schools technicians are expected to support.

COMPONENT 10 – ISSUES REGARDING WINDOW OF ADMINISTRATION, TEST SECURITY, AND NEED OF A BACKUP SYSTEM FOR STATE, DISTRICT, SCHOOL, AND CLASSROOM PURPOSES

This component of the study addresses two distinct, but related components: overall test security and the issues surrounding testing windows and the need for well-defined risk management plans that address any issues resulting in system downtime.

Testing Security

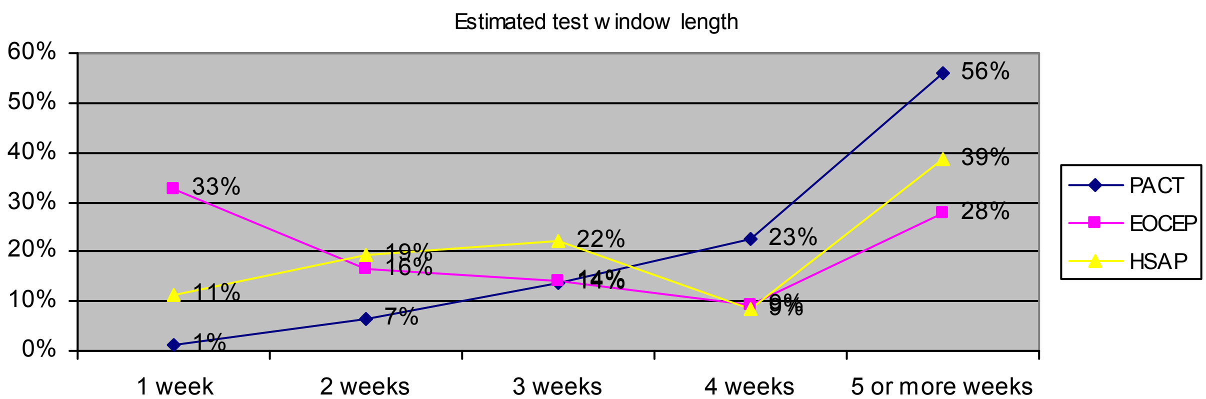
When evaluating a move to an electronic testing program, especially in programs where some type of stakes are attached to the result of the test, overall security of the program must be a driving factor. A comprehensive testing security plan is a key component in any assessment program, and electronic delivery models are no different.

This study will review key elements of a high quality test security plan and how they differ between paper/pencil and electronic assessment programs.

ITEM SECURITY AND RISK OF EXPOSURE

A large-scale assessment program is often only as good as the test items associated with the program. The quality of the test items and their alignment to the testing standards ultimately affect how the assessment program meets its defined goals. The development of test items is often one of the more costly portions of a large-scale assessment program. Consequently, a great deal of value is placed on keeping test items secure in the present and future.

Electronic delivery of large-scale assessment programs often results in the lengthening of test windows to accommodate the number of students that must be tested. Currently the vast majority of computers that schools have available to be used for testing reside in computer labs. When surveyed, 92% of South Carolina educators said, should electronic testing take place, the primary location for student testing would be in school computer labs. This is consistent with experiences in other state programs that have moved to electronic testing. In order to accommodate the number of students that must be cycled through computer labs to meet the testing needs of the state, the length of the testing window may need to be increased. However, increasing the length of the testing means that secure test items are now exposed for a longer period of time. This study asked test administrators to estimate the length of the testing window they would require if the state's current assessment programs were moved to an electronic delivery platform. The results are shown in the figure below:



Technology and process changes assist in mitigating the additional risk of item and test exposure during these longer testing windows. Despite longer testing windows, a fully implemented electronic testing program may actually reduce the total amount of time items are potentially exposed relative to paper/pencil programs. In an assessment program that is fully or partially implemented using paper/pencil, the total amount of time items could be exposed begins the moment the test booklets arrive at the testing site. States have programs to safeguard paper tests, and these include manual steps that depend on humans following the correct steps. Because paper can easily be copied with little or no way of tracking, the items remain exposed 24 hours a day, 7 days a week throughout the length of the testing window or until the test booklets are returned to the vendor for processing. In a fully electronic testing model, the test items are only exposed while students are physically at the computer taking the test.

To illustrate this point, assume the delivery of a test that takes students approximately one hour to complete. In the paper/pencil delivery of the test, further assume a one week (Monday through Friday) testing window and that the testing materials arrive one business day ahead of the testing window and are shipped back one day after the close of the testing window. This would translate into 264 hours or 11 days of total item exposure time. For the electronic testing example, assume a three week testing window and seven hours per day when students are taking the test. This would result in 105 hours of item exposure. In this example, total item exposure time is reduced by approximately 60 percent.

The most secure electronic testing solutions encrypt test items and responses as they are being transmitted or for the time period they exist on the computers or servers. Also, a quality high-stakes testing solution should utilize software that prevents the printing or electronic copying of any secure test item. If an item or test does need to be printed, an electronic record of the time, location, and responsible party should be captured for future reference. These features make the copying of test items for malicious purposes much more difficult and provide an electronic "paper trail" to assist in investigation.

Security of the Test Delivery System

The security of the test delivery system in electronic testing should address two key factors:

- preventing the computer from being utilized in a way other than intended by the test creators to complete the test and
- securing the test and related data from outside parties or hackers.

Today's software is designed to ease the burden on the user and to provide access to tools that can be used for a variety of purposes. Examples of these ease-of-use features include access to the Internet, search engines, calculators, email and instant messaging. While all of these are useful tools, they can work against the need to create both a highly-standardized and highly-secure large-scale testing environment. The Internet could be used by students to find answers to questions. Calculators may provide unfair advantages to some students if they are not utilized in a standard way. Email or instant messaging could be used by a student to communicate with an outside party to gain an unfair advantage. The software used to deliver large scale, standardized tests must address these factors when being implemented.

As discussed in Component 6, most software utilized to deliver tests electronically today follows one of two basic delivery mechanisms: browser-based or smart client. Each delivery mechanism has security considerations.

The browser-based solution utilizes a standard web browser such as Microsoft Internet Explorer, Firefox, or Safari. In this architecture, the user logs onto a secure testing site with a username and password much like he/she would any other site. The testing program is encrypted through SSL (secure socket layer) or another standard encryption protocol. However, this architecture may not block or prevent access to other system tools or Internet access. For this reason, this architecture is typically preferred for low-stakes or formative assessments or where test content is not secure. Some testing vendors have extended this architecture to block commonly used key sequences to access external programs in an attempt to make this architecture more "secure." However, this should not be mistaken for a truly secure browser.

The second delivery mechanism is commonly referred to as a secure browser or "smart client." In this method, software is loaded on each computer that will be used for testing. When started, this software takes control of the user's machine and can prevent access to any unauthorized web sites, system tools, or communications software. This software is proprietary to the testing vendor and is the responsibility of the vendor to create, test, and maintain. This is the preferred approach for high-stakes assessment programs where standardization and security are key because it is currently the only method to ensure that test takers cannot access external programs or systems in an attempt to gain an advantage.

The security of the testing system against potential unauthorized intrusion from "hackers" or other individuals who may use the information for unintended or malicious purposes is shared by the state and its testing vendors. The state's responsibilities in this area are threefold:

1. The state must clearly communicate the security expectations of the program. Security needs can vary greatly from one assessment program to the next. It is incumbent upon the state to work closely with the testing vendor to define and approve all levels of security procedures in the testing program.
2. The state should take a "trust but verify" stance with the testing vendor. Once the security expectations of the program are communicated, the state has a duty to verify that the expectations are being followed. It is strongly recommended the state ask for, and carefully review, the security plans of the assessment vendors. If necessary, implement an independent security audit of the testing vendors to validate that both the requirements of the state and industry best practices are being followed. A more detailed discussion of the security audit is addressed in the Security Plans & Audit subsection.
3. The state and its testing sites also maintain a responsibility for the security of their own network and attached devices as well. This includes implementing appropriate firewall, anti-spyware, and anti-virus programs on all network components, and taking steps to make sure they are up-to-date. If the testing site is utilizing any type

of wireless networking technologies, it would also be incumbent upon the testing site to apply all appropriate encryption algorithms to prevent unauthorized access.

The vendor's responsibilities are to:

1. Design and implement standard security features and best practices in its test delivery solution.
2. Work with the state department of education to understand and implement the state's specific security requirements.

The major message in this section is that there are a variety of security considerations, and the state and the vendor must work together to ensure the state's needs are being met.

SECURITY OF THE STUDENT REGISTRATION SYSTEM

The testing vendor accepts primary responsibility for the security of the student registration system. It is incumbent upon the state to clearly define the security needs of the program and to verify their implementation. The Family Educational Rights and Privacy Act (FERPA) places certain restriction upon what student information schools may disclose without parental consent. Testing vendors must know and adhere to the regulations specified in FERPA because they will have knowledge of student demographic information as well as data directly pertaining to the student's assessment and educational record. This is the same whether the assessment program is paper-based or electronic.

SECURITY PLANS & AUDITS

Assessment programs present a multitude of risks and complications for states as well as for the testing vendors that support the programs. Changes in technology, new or changing state and federal legislation, and procedural changes to the assessment program itself all impact the overall security framework of the assessment program. It is incumbent upon the state to validate that their testing vendor has appropriate security designs and safeguards in place. One approach to validating this is for the state to request formal security plans from the testing vendors. These plans should highlight all of the steps that the testing vendor is taking to ensure the overall security of the program. These plans should be reviewed and approved by knowledgeable parties at the state level and any vendor shortcomings should promptly be addressed. Additionally, the state should be aware that it is possible for vendors to apply for and receive specific security certifications. These certifications can be an indication to the state that the vendor is implementing current industry standards for security.

If the state desires an even greater level of security from its assessment vendors, it may request an independent security audit be performed. An audit may take various forms, but the intent is consistent; to ensure the documented security policies of the testing vendor meet industry standards and the needs of the assessment program.

It is important to note that audits such as those described above often come with a significant cost, anywhere from $50,000 to $150,000 depending on the complexity of the program and the level of detail desired. This study was not able to find any state currently asking this of its assessment vendor. However, it is an option that can be considered if the circumstances warrant.

BACKUP AND FAILOVER PLANS

While electronic test delivery in considerably more stable, secure, and reliable than it was only a few years ago, the nature of using a computer and the Internet to deliver assessments still increases the potential for issues to arise. Computers break. Servers or network devices may fail unexpectedly. Communication or power lines could be cut or downed by a storm. There is always the potential for some type of hardware failure. This reality makes it incumbent on the state to work closely with the chosen testing vendor to develop risk mitigation strategies and failover plans that can be implemented quickly in the event it becomes necessary.

In the broadest sense, most hardware or infrastructure failures fall into two categories. The first is a hardware failure or other performance related issue with the testing vendor. Examples of this type of failure could include the vendor's servers going down, or the vendor's solution is not capable of meeting the performance requirements of the state. When an issue occurs with the testing vendor, the consequences can be severe, as it typically affects most if not all of the testing sites in the state at one time. When this type of failure happens, it is vital that the state and vendor have agreed to predefined courses of action to minimize the impact and number of testing days lost. Some potential courses of action include:

- implementation of backup hardware by the vendor (this should already be in place);
- packaging and delivery of preprinted paper tests to affected testing sites; and
- extensions to the testing window in the event the issue cannot be corrected in a timely manner.

The second type of hardware failure occurs at a single or perhaps localized number of testing sites. This may include situations such as power outages, Internet connectivity issues, or other issues. They may be isolated to a single district, single school, or even a single lab within a school. These issues may not have great impact on the testing program as a whole, but can still have a significant impact on the ability of the affected site to complete testing. Scheduling of computer labs during the assessment window can be a difficult task under the best of circumstances and any disruptions to that schedule may impact other planned learning activities. Computerized Assessments & Learning, the group who delivers electronic NCLB tests for the state of Kansas, suggests the following to administrators when scheduling lab time: "Plan for at one 'snow day' for each five days of testing you need to complete. A 'snow day' is a backup day to help you work around any type of scheduling disruption."

States should incorporate a failover plan, or risk management plan, as part of their implementation of an electronic testing program. This plan should clearly identify the areas of risk and the programs in place to mitigate or respond to those risks.

A critical component of any type of significant backup strategy should be a well thought-out communication plan. If the state must decide to implement a backup plan such as the distribution of preprinted paper tests, it will be critical to get this information out to the testing sites as quickly as possible. The state should work closely with its testing vendors to determine an appropriate communication plan for each of the backup and failover strategies it develops. In situations like this, there will be a high level of discomfort in the field already. Good communication can go a long way to reducing that level of discomfort.

PLACEMENT OF COMPUTER IN LABS

A sometimes-forgotten area relating to overall test security is the use of computer labs themselves in the assessment process. In most schools today, limited space means that computer labs are configured with computers in close proximity to one another and often have little or no desk writing space for the student. Also, computer monitors are placed at an angle or on risers to allow instructors easy viewing of student work. While this setup works well for instructional purposes, it is not ideal for conducting standardized assessments. The easy viewing of computer monitors along with their close proximity to one another increases the potential for cheating and compromising the test. States have developed a number of courses of action to mitigate these risks ranging from the simple to “high tech.” One approach is to use the computer and software to randomize the item sequence on the tests. Another alternative would be to use multiple test forms and not allow students with the same form to sit next to each other. This significantly reduces the likelihood that two students seated next to each other will see the same item at the same time, but has some offsetting test design considerations. Another approach could be the use of privacy filters on all computer monitors used for testing purposes. These filters prevent someone sitting at an angle to the right or left of the monitor from viewing what is currently display on the screen.

Many schools and testing locations have also fashioned temporary partitions using everything from wood to fabric to manila file folders as a means of preventing students from “looking next door.” It seems the best method for addressing these types of configuration issues is for the testing vendor to work closely with the testing site to understand the issue and offer suggested courses of action based on experience in other locations. It is also important for the state to establish communication channels that allow test administrators to communicate easily with one another so they may share what has worked well with their colleagues.

Summary

This section discussed the security implications of moving from a paper/pencil-delivered assessment program to an electronically delivered program. The security of an assessment program can be large and multi-faceted. When considering a move to electronic delivery of an assessment program, there are several factors to consider:

- Item security and exposure
- Security of the test delivery system
- Security of the student registration system
- Security plans and audits
- Backup and failover plans and strategies
- Placement of computers in labs

In order to maximize the overall security of the assessment program, this study recommends the following courses of action:

- The state should work closely with its testing vendors to define the overall security requirements for the testing program. These requirements establish the baseline for testing and assessment.
- The state should take a "trust but verify" stance with assessment vendors. Careful review and approval of vendor security plans are a must. The state may also look for vendors to meet nationally accepted security certifications. If necessary, the state may request an independent security audit of the assessment vendors.
- The state should work closely with vendors to establish backup and failover plans for as many situations as possible. These plans may vary in the cost and complexity in

relation to the potential risk. However, for each plan, a clear communication approach should be developed and ready to implement.

- The state should establish communication channels that allow test administrators and technology coordinators to easily communicate with one another. By facilitating communication between parties in the field, the state can address issues more quickly, share lessons learned, and reduce discomfort levels when issues do arise.

COMPONENT 11 – DELIVERY OF RESULTS FOR SCHOOLS AND STUDENTS AND THE ABILITY TO PROVIDE INSTRUCTIONALLY INFORMATIVE RESULTS TO DISTRICTS, SCHOOLS, TEACHERS, AND PARENTS

As South Carolina considers moving toward more timely, instructionally informative reports from statewide testing initiatives, several aspects of other state testing programs are presented below in order to provide context for South Carolina's decision-making process. Information regarding testing windows, report turnaround times, and testing resources for all 50 states has been compiled, whether the testing programs are paper/pencil-based, CBT/CAT, or some combination. See Table 11.1 at the end of Component 11.

Additionally, this section concludes with a discussion of options for providing Computerized Testing Score Reports that are linked to instruction.

Summary of Findings From Other States

OVERVIEW OF TEST DESIGNS

Thirty-three states offer some combination of multiple-choice tests with an open-ended/constructed-response/short-answer/extended response component, excluding writing as a separate test. Thirty-five states, including South Carolina, assess writing in at least one grade, with one state having plans to add writing and two other states "officially" assessing writing at the local level. While the scoring of open-ended and/or writing assessments generally increases the time between testing (whether via computer or paper/pencil) and reporting, most states have determined that some sort of assessment that goes beyond multiple-choice is worth the cost and the slower reporting.

REPORTS

The vast majority of states provide some sort of reporting at a strand/standard/objective/subcategory level, whether by raw scores, percent correct, or a Standard Performance Index. The minimum number of items/points for reporting at the standard level range from 1–12, with a minimum of 6-8 being the most common. (The lower minimums most often apply only to grade 3 and only to one or two strands/standards. As one goes up the grades, the minimum number of items used for strand/standard-level reporting is higher.) South Carolina is one of a handful of states that generally report scores only at the overall content/subject level.

TESTING WINDOWS

There is a great variety of testing windows among the states. Many states have a single day for testing each subject, while others have testing windows as long at 3 to 4 weeks or more. (e.g., Iowa, using the ITBS, allows testing between early September and late April). Understandably, testing windows are often longer for CBT/CAT testing than paper/pencil (e.g., 4-6 weeks in Idaho or Kansas, vs. a single day for paper/pencil testing of a grade and subject in South Carolina or Texas). However, many paper/pencil-based testing programs also have extended testing windows.

The time of year testing is conducted also varies—at least seven states (Indiana, Michigan, Rhode Island, Vermont, Wisconsin, New Hampshire, and North Dakota) test in the fall rather than the spring. Separate writing tests are often given earlier in the year than the multiple-choice-only tests or multiple-choice/open-ended tests to allow additional time for hand scoring. At least nine of the 35 states that assess writing separately do this, with an additional five states assessing English language arts/reading/writing in some combination earlier in the year. For a discussion of the feasibility of South Carolina moving its writing tests to earlier in the year, please see Component 5.

REPORT TURNAROUNDS

The time between testing and reporting also varies widely by state and testing program. Computerized testing tends, not surprisingly, to have a faster turnaround than paper/pencil-based tests. For CBT/CAT, preliminary scores can be immediate/almost immediate (e.g., Idaho, Oregon—multiple choice only), or, as in the case of Virginia, which does post-equating of certain CBT forms, there is a target of less than two weeks for reporting (pre-equated forms are reported immediately). Other computerized tests that do not provide immediate reporting include: Florida (high school retesters, multiple-choice only, all data available in six weeks); Texas (reports in two weeks); Mississippi (end-of-course retesters' pass/fail rosters available in three weeks; Wyoming (multiple-choice, second week of testing window; results including short-constructed response available four weeks after close of testing window); and West Virginia (writing scored by Artificial Intelligence reported in 60 days).

For paper/pencil tests (the vast majority of NCLB-mandated tests), the turnaround time generally ranges between 6 weeks and 16 weeks. However, multiple-choice-only tests *may* be reported more quickly (e.g., Utah reports raw scores back to a district four days after receipt of materials, although this is unusual). If open-ended responses are included, the turnaround time may be longer, depending on the state. Again, in nine states writing as a separate test is administered earlier in the year than other subjects to allow additional time for scoring. Additionally, at least five states administer some or all English language arts, reading, or combined reading and writing tests earlier in the year than other content areas.

Finally, certain high-stakes paper/pencil tests, such as graduation exams or "gateway" tests for lower grades are often placed on a fast-track for scoring.

Test Blueprints/Test Specifications

With the implementation of the No Child Left Behind Act, all states are now providing test blueprints or test specifications. Therefore, there is not a state-by-state accounting of this aspect.

Additional Testing Information Provided by States (e.g., Sample Items, Released Tests)

Information collected from states indicate that most states provide, at a minimum, sample items, and at the maximum, released test forms for every grade and subject, every other year (i.e., Texas). Some states provide more than one resource, and the terminology may vary from state to state (e.g., one state may label a resource as a practice test, while another may label the same sort of resource as a sample test). An accounting from available information shows that:

- 4 states provide sample tests
- 17 states provide practice tests
- 6 states provide released tests
- 26 states provide released items
- 19 states provide sample items

The financial implications of releasing operational items/forms cannot be determined for a number of reasons:

- If a state has one contract with a single vendor, those contracts often do not break out item development costs.
- If a state contracts with one vendor solely for item development, with a specific per-item cost, those costs do not include another vendor's cost for field-test form development (if stand-alone field tests) or creation of multiple forms for embedding field test items. Also not included are costs for pulling field-test samples, data collection and analysis, etc.
- "Hidden" state-level staffing time/costs for additional review committees, editing rounds, data review, etc. cannot be determined.

In short, the more item development and associated costs needed for releasing items and/or forms, the higher the explicit and implicit costs to a state. Also, each state's potential pool of students for field-test sampling and a state's use of linking or anchor items are additional considerations when determining the release of items and/or forms.

STATE SCORE REPORTING THAT PROVIDES ACTIVITIES AND/OR INSTRUCTIONAL LINKS FOR PARENTS AND/OR TEACHERS

The research shows that there is limited but apparently growing state-level interest in using student score reports from large-scale assessments as a vehicle for linking student-specific performance with activities/resources intended to improve student learning.

From a research standpoint, this interest is reflected in a 2006 whitepaper, "Developing Score Reports for Statewide Assessments that are Valued and Used: Feedback from K–12 Stakeholders," published by Pearson Educational Measurement (Trout & Hyde 2006). The authors describe a two-phased focus group study involving various educational stakeholder groups that investigated "… ways to maximize the fit between test results and the educational contexts in which those data are used" (p.1). Their primary research goal was to "…determine how users/stakeholders use statewide assessment results and to determine what score-reporting information adds the most value and allows users/stakeholders to take appropriate next steps" (p.1). One of their more interesting findings involved parents' feedback on mockups of hard copy reports that were further supported by an associated website. Parents indicated they liked the links to learning activities and resources, including online interactive and printable worksheet types of resources, particularly for younger students (p.16).

Research on educator interest in such "enhanced reports" is limited. However, in 2007 Data Recognition Corporation commissioned Eduventures to conduct a sampling survey of state- and district-level staff on this topic. The results were reported in an unpublished presentation titled, "Perceptions of Enhanced Reporting for State Assessments." At best, the results from a limited sampling of state and district administrators were mixed, with some administrators expressing interest in enhanced reports, others noting little interest in such reports from their constituents, and a number indicating that they would be adopting a "wait and see" posture with regard to possibly providing such reports.

With respect to current implementation of such "enhanced reports," only a few states are providing reports that link assessment with instruction/activities for parents. A notable example is American Institute of Research's reporting system for Ohio. Student reports in this state provide a "Next Steps" section, which includes activities that parents can do with their children (e.g., in Reading, under the standard Reading Process — "Have your child read and summarize a challenging story, article or piece of nonfiction. Ask your child to tell you what she thinks about the text and to support this opinion with details."). Typically, each standard has an associated strategy based on student performance.

Also, the Grow Network, through a combination of hard copy student reports and an associated website, is providing parents in Pennsylvania, New York, and Nevada with links to home activities. These activities, found at www.growparent.com, are not necessarily tied to student-specific results. Rather, they are general collections of activities, typically broken down by grade level and difficulty (Beginning, Intermediate and Advanced) in both English language arts and mathematics. The math section draws largely from the University of California, Berkeley, Family Math program.

A slightly different approach to linking student-specific results to activities is the Grow Network/McGraw-Hill's MyGuide Program. This program provides "personalized study guides" for high school students in two states, Texas and Arizona, and the Grow Network has recently announced its expansion to California. According to the Grow Network website, students who have not passed at least one portion of their state's high school exit exam may visit a state-specific website and enter their scores (overall and by objective/strand), and a customized study guide is generated, including guided practice and tutorials. (Texas also makes these guides available in hard copy to students in Grades 9 and 10.) Additionally, resources for teachers, tutors, and parents are provided to assist students in using their "personalized study guides." To date, this initiative is only available to high school students; and the Grow Network/CTB McGraw-Hill reports that, "the National Governors Association has showcased the personalized study guides as one of the top ten initiatives being implemented to support high school reform efforts."

Beyond the examples noted above, it appears presently that most efforts to link instructional activities/resources with statewide testing results remain the purview of district/schools. Still, the ever-growing desire for data-driven decisions/instruction, suggests an increasing state-level interest in linking student results with resources for both parents and educators.

FORMATIVE ASSESSMENTS

This is an area of statewide initiatives in particular flux. Available information indicates states have generally taken one of several tacks:

- A list of state-approved/reviewed formative assessments from which districts may choose (e.g., Pennsylvania, South Carolina)
- State-developed formative assessment item banks for NCLB tests (e.g., Arizona, in conjunction with Arizona State University)
- Online state-developed, sanctioned, or adopted formative assessments for NCLB tests (e.g., Alaska, Kansas, South Dakota—the latter both CBT and CAT)
- Online formative (or more specifically, diagnostic) tests for limited subjects/grades (e.g., Texas' Math and Science Diagnostic Systems, grades 4-8; Virginia's Algebra readiness assessment for middle school students; Michigan's planned end-of-course formative assessments).
- No state involvement in formative assessments—district-level choice

Further information regarding effective and aligned formative assessments should be forthcoming from the Council of Chief State School Officers' (CCSSO) Formative Assessments in a Comprehensive Assessment System Study Group. This group is planning "papers and tools to provide timely assistance to States in their work," as well as the creation of a guide for states. For more information see the References and Links section at the end of Component 11.

Overall, there has been an increased demand for standardized, state-standards-aligned formative assessments, as the high-stakes nature of NCLB testing has permeated education.

There seems no one path that states have taken, whether due to local control issues, funding availability, coordination among state divisions and/or districts, or staffing resources.

In the best of all possible worlds, state-sanctioned/developed formative assessments, related instructional resources, and data from summative assessments would work in a seamless day-to-day instructional and data-driven environment. However, the funding, leadership, coordination, and commitment to such an endeavor has yet, apparently, to be *fully* realized by any state.

Please see also the recommendations for South Carolina formative assessments in Component 14.

DISCUSSION OF THE POTENTIAL FOR SOUTH CAROLINA COMPUTER-BASED TESTING SCORE REPORTS THAT ARE BOTH TIMELY AND USEFUL

Reporting—Timely, Useful Results

First and foremost, timely, useful reports depend on having a valid, reliable[1] test to report. Given that starting point, the most meaningful metric to report scores is a *scale score* that has been constructed to eliminate irrelevant characteristics of tests. For this purpose, *irrelevant* characteristics include the test length and the attributes of the specific items administered. The test must be long enough to achieve the desired degree of precision (reliability) and the specific items must be appropriate to the construct being assessed (validity). If those are givens, one need not know any more about the test to interpret the scale score.

Scale scores for students can be compared directly to scale scores based on other selections of items, to the locations of items, and to the performance standards. Regardless of how the results are ultimately reported and how many items are involved, scale scores will be the underlying metric that makes meaningful analysis and interpretation possible.

[1] When considering the results for one student, *reliable* means a small standard error.

Student-level, diagnostic reports for computer-administered, multiple-choice tests could be delivered to the teacher's computer screen before the student can logoff from the computer and walk to the teacher's desk. If the items have been *calibrated*[2] and *performance standards*[3] established prior to the assessment, then, at a minimum, the following information could be reported:

1. Scale score,
2. Confidence interval for the scale score,
3. Performance level,
4. Probability that the student might truly be in a higher or lower level,
5. Evaluation of the student's performance for each content standard, item type, passage, or any relevant cluster of items,
6. Disclaimer reminding the user that these results are preliminary.

Bullets 1 through 4 facilitate summary and analysis, hence are more for accountability than for altering instruction. Number 5 is for the teachers and does not require pre-calibration or performance standards. It should include individualized text that highlights the conclusions that are warranted and avoids ones that are not. This needs to be more than using the scores in complete sentences. For example, statements like,

"Abby has a scale score of 1234 with a probable range of 1195 to 1273, which places her in the *Proficient* Performance Level"

[2] In the narrowest sense, *calibrated* means that the item difficulties parameters have been estimated. As intended here, it means, in addition, that the items have been thoroughly reviewed and approved, that they functioned appropriately in a tryout, and that the estimated item difficulties are all on a common scale (i.e., *pre-equated*.)

[3] *Performance standards* are the quantitative definition of the performance levels. They might say, for example, that a scale score of 1200 is required to be classified as *Proficient*. These standards are expressed in the same metric as the scale scores and the item difficulty estimates.

add very little value beyond what the typical user could have gained from a score summary table for Abby. Ideally, it will draw insights out of the data that are beyond the obvious conclusions that any parent might draw. It might be more informative to say something like,

"Abby's score of 1234 places her solidly in the *Proficient* Performance Level with less than 10% chance that she would be placed in *Basic* if tested again."[4]

The report provided to the teacher or parent need not (and definitely should not) include anything about *logit item residuals, likelihood ratio tests,* or *weighted between group mean squares*. The data shown for each item cluster can be as simple as mastery-type scores, e.g., the number correct out of the number possible. Based on the tenuous assumption that the items on the test are a representative sampling of all the possible items for the content domain, the mastery statement will tell the teacher if the student can answer items on that topic or, perhaps an estimate of percentage of items in the domain the student can answer. Estimating this type of probability from a small sample of items is more problematic than estimating a scale score.

The analysis should do more to help extract a meaningful and defensible interpretation of the results.

"In *Number Sense*, Abby answered five of eight items correctly. Based on her scale score of 1234, we expected her to answer six correctly. One of the items she missed was very easy for other students, so this may indicate a topic she needs to review or it may have been a careless error."

This simple graphic (Figure 11.1) is based on the underlying measurement continuum although that metric is not explicitly shown[5]. The chart shows the relative location of the number correct scores (solid line with 9 diamonds as well as the integers 0 to 8 at the top of the figure); it shows Abby's mastery score (large star under the score of 5), the expected score on

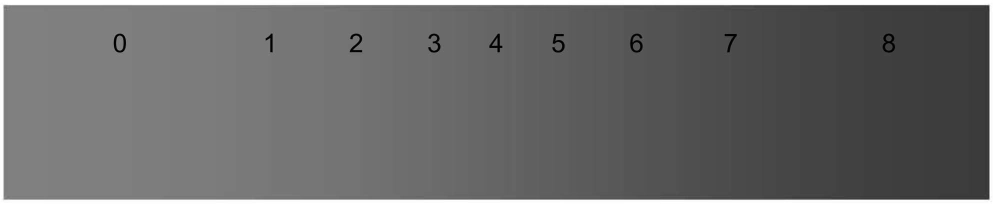
---

[4] These sample statements were written by a psychometrician. Determining the most informative statements and effective wording should involve educators from all levels.

[5] The scale score metric could be shown as well if that has been established and the items pre-calibrated, but it really is not needed.

this standard (small star under the 6), the relative difficulties of items answered correctly (five small squares above the line,) and the relative difficulties of items answered incorrectly (three squares below the line.) Producing this graphic, intended for the teacher, does not require the items to have been pre-calibrated or performance standards established.

Figure 11.1: Sample Display Showing 5 Correct Responses to 8 Items.



This chart does not include any indication of the standard error[6], measurement error, confidence interval, or probable range. It does show how far the score on this standard was from the expectation and leaves it to the teacher to decide if 5 is different than 6.

The sample chart can be included in a printed report, with an appropriate explanation. When the results are presented on computer, the explanation could be in the form of a tutorial. Context-specific help could be available by rolling the pointer across any element of the display. Details for specific items could also be obtained by pointing to the square for that item. This could be all the content standard information associated with the item; it could be a key-word descriptor of the item that describes the problem without giving away the item; or, if the items have been released, it could be an actual image of the item[7].

All computerized testing contractors have the capability to provide immediate feedback to the examinee. Most systems, however, were not developed in the context of large scale, accountability assessment. Most are either designed for formative assessment or licensing and certification. In most cases, these systems will require some modification to provide the reports

[6] The standard error is approximately two raw score points.

at the level of sophistication suggested here. The basic number-correct reports should be readily available and may satisfy teacher expectations, although they do not exploit well the available computer or psychometric technology and they are completely inconsistent with CAT.

Constructed Response / Extended Response Tasks

CR tasks would require slightly different reporting because they do not have dichotomous scores. Rather than showing a single location for each item, there is a location for each possible score on the task. The diamond line in the figure above could represent a single 8-point CR as well as it does an 8 MC subtest. Because of the underlying scale of measurement, the CR/ER results can be displayed along side the MC data and the student's results.

Number of Items Required for a Reliable Score

Whether administered and reported via paper or computer, and no matter what metric is used for reporting, a major issue is the number of items needed for a standard to justify reporting student performance at the standard level. There is a constant conflict between the desire to provide as much detailed diagnostic information as possible and the need to keep testing time to a minimum.

Typical recommendations are that eight to ten items are necessary to report a score, although many programs use fewer. In general, precision, defined as the measurement error, is proportional to the inverse of the square root of the number of items. Consequently, a score based on eight items would be expected to have twice the measurement error as a score based on 32 items. In terms that are more familiar but less appropriate for a standards-based test, the reliability for an eight-item scale is unlikely to be as high as 0.4; with 32 items, the reliability should be above 0.8.

---

[7] Teachers would be delighted to have the specific item information. Providing it creates the risk that instruction would then focus on answering that specific item rather than correcting any underlying misunderstanding.

This discussion does not depend on whether the results are reported in scale score or raw score metrics. The raw score metric is more familiar but has the same ambiguity for interpretation that makes reporting raw scores problematic for tests of any length. It merely serves to disguise the uncertainty associated with a short test.

Dimensions of Diagnostic Reports

There is a delicate balance to be struck for assessments that are tightly linked to content standards and that report status by performance level. To report the status of a student based on a single summary score implies the underlying trait is unidimensional, analogous to distance or weight. This justifies such statements as, “Abby is proficient in Mathematics.”

Although every item is aligned with a specific content standard, the assumption is they all add up to one thing. Items typically vary in difficulty and so are spread along the continuum but they are all on the same continuum. One could, in principle, make inferences about a student’s mastery of one content standard based on how that student performed on other content standards. This is the practical implication of a unidimensional trait.

On the other hand, Luecht et al. argue that useful diagnostic scoring requires reliable, multidimensional measurement information (Luecht, Gierl, Tan, and Huff, 2006). One can usually confirm with factor analysis or other techniques the presence of factors associated with the content standards although the factors are typically weak. The presence of multiple factors would imply the scores should not be summed to determine a single status for a student but should be reported standard by standard.

Summary

The practical considerations of diagnostic reports can be summarized as follows:

1. Number-correct scores are not the appropriate metric to report student performance. Scale scores eliminate the ambiguities associated with test length and item difficulty, which makes possible meaningful comparisons across standards, across forms, across years, and, potentially, across grades.
2. A student's performance on a standards-based assessment can usually be summarized by the scale score derived from the total score across all standards.
3. For most students, instruction and learning follow a common sequence so that an advanced standard will not be mastered until the preliminary standards are mastered. Diagnostic reporting relies on knowing the normal path and how far along the student has progressed.
4. This simple model is not a complete description of some students who, for some reason, do not follow the normal path. These students may have unusual abilities or unusual out-of-school experience. Their diagnostic profiles may show surprising mastery of advanced standards that are not predicted by their performance on the preliminaries or may show surprising gaps in the preliminaries that did not prevent success on the advanced standards.
5. The number of items, or points, that contribute to the score for a standard is a trade-off between precision and testing time. There is no definitive answer to the number of items necessary for a *reliable* score but eight to ten items are typical values cited.

The information in Table 11.1 has been collected from multiple sources, including Internet searches of state websites, e-mail correspondence to various state staff, and follow-up telephone calls to state departments. This information is as up-to-date and correct as possible, understanding that state testing programs are often in flux. Also, the research has focused on NCLB-required (3–8) tests. When additional information has been readily available, that has been included as well. The section does not include data on all aspects of state testing programs, such as the so-called 1% and 2% alternate assessments, English Language Proficiency tests, or multiple testing dates for high school graduation or end-of course examinations.

Table 11.1: Fifty-State Reporting and Testing Information						
State	**Test**	**Reporting Categories + Minimum # Items for Strand Reporting—Raw Scores/Scaled Scores % Correct, If Applicable**	**Multiple-Choice/ Open-ended (+ Writing)**	**Sample Items, Released Tests, Practice Tests, Released Items**	**Test Dates/ Window**	**Score Reporting**
Alabama	ARMT (Reading Math) Grades 3–8	Minimum 6 for Raw Score reporting at the standard-level	MC/OE	Sample Items	4/2–4/13	10 weeks electronic/ 11–12 weeks paper
	AHSGE		MC		4/5–4/9	3 weeks electronic/6 weeks paper
Alaska	SBA (Grades 3–9)	Minimum 8 points, reported as both Raw Scores and Scale Scores	MC/OE + writing	Sample Items/ Practice tests	4/2–4/16	4 weeks electronic/6 weeks paper
	Grade 10 SBA-HSGQE & HSGQE Retest	Minimum 8 points, reported as both Raw Scores and Scale Scores	MC/OE + writing	Sample Items/ Practice tests	4/3–4/5	Retest = 2 1/2 weeks, electronic/6 weeks paper; Others = 5 weeks electronic/6 weeks paper
Arizona	AIMS DPA (Grades 3–8)	Minimum 5 items, reported as Raw Scores + % correct	MC+ writing	Sample Items/ Practice tests	4/7–4/19	6 weeks paper
	AIMS High School	Minimum 5 items, reported as Raw Scores + % correct	MC + writing	Sample Items/Sample Tests	Reading/Writing, 2/27–3/7 Math, 4/11–4/17	6 weeks paper
Arkansas	ACTAP (Grades 3–8)	Minimum 6 items, reported as Raw Scores	MC/OE	Released Items	4/17–4/20	9 weeks paper

Table 11.1: Fifty-State Reporting and Testing Information						
State	**Test**	**Reporting Categories + Minimum # Items for Strand Reporting—Raw Scores/Scaled Scores % Correct, If Applicable**	**Multiple-Choice/ Open-ended (+ Writing)**	**Sample Items, Released Tests, Practice Tests, Released Items**	**Test Dates/ Window**	**Score Reporting**
California	STAR Grades 3–8	6 items (grade 2), reported as % correct	MC + writing	Released Items	21-day window—the 10 days on either side of date 85% of instruction completed. Writing—March 6	07/13–8/8; 16 weeks for writing
	CHSEE (High school proficiency exam)	Minimum of 4 items, reported as Raw Scores	MC + writing	Released Items	Various	8 weeks
Colorado	CSAP Grades (3–10)	9 items minimum for reporting (based on grade 3 item maps 2005). Above/Below Proficient for Standard/Sub content levels reported by for ISRs; Scale scores for School reports, by each level, by student.	MC/OE + writing	Released Items/ Item Maps	Grade 3 Reading 02/13–02/24 All other 3/13–4/14	Grade 3 electronic 4 weeks/paper 9 1/2 weeks. All other 14 weeks electronic/paper 16 weeks
Connecticut	CMT (Grades 3–8)	4 items (grade 3 math), Raw Scores + mastery indicator	MC/OE + writing	Sample Items/ Practice tests	3/5–3/30 Writing fixed for 3/6	12 weeks electronic/ 14 weeks paper

Table 11.1: Fifty-State Reporting and Testing Information						
State	**Test**	**Reporting Categories + Minimum # Items for Strand Reporting—Raw Scores/Scaled Scores % Correct, If Applicable**	**Multiple-Choice/ Open-ended (+ Writing)**	**Sample Items, Released Tests, Practice Tests, Released Items**	**Test Dates/ Window**	**Score Reporting**
Connecticut (continued)	CAPT (10)	12 items, Raw Scores	MC/OE + writing	Released Items	3/5–3/30 various writing/lit. response subtests, fixed for 03/06–03/08	
Delaware	DSTP (Grades 3–8, HS)	Reading and Math as content area scale scores/HS Science and Social Studies scale scores + raw/strand. Minimum # 11 points/strand	MC/OE + writing (Grades 3, 5, 8, and 10)	Sample Items	3/16–3/28 Grades 2–10	June ELA and Math—July Science and Soc Studies = 10–14 weeks
Florida	FCAT	Minimum 2 raw score points, percentages only reported	MC/OE + writing	Sample Tests/ Released Items	Writing, 2/6–2/9; Reading, Math, Science, 2/26–3/9	High Stakes (grade 3, writing, graduation retakers), late April = approx. 7 weeks (6 weeks for online retesters); all others, late May = approx. 11 weeks
Georgia	GHSGT (High School)	11 items minimum for strand reporting	MC	Sample Items	3/20–3/31 (Spring admin only)	Electronic—4 1/2 weeks, Paper—5 1/2 weeks
	CRCT (Grades 1–8)	Minimum of 9 points—raw scores reported	MC	Released Tests/ Sample Items	4/3-5/3	3–7 weeks
		Writing tests at Grades 3, 5, 8, 11	Extended Writing		Grade 5, 1/17–1/20	11 weeks

Table 11.1: Fifty-State Reporting and Testing Information						
State	**Test**	**Reporting Categories + Minimum # Items for Strand Reporting—Raw Scores/Scaled Scores % Correct, If Applicable**	**Multiple-Choice/ Open-ended (+ Writing)**	**Sample Items, Released Tests, Practice Tests, Released Items**	**Test Dates/ Window**	**Score Reporting**
Hawaii	HAS (Grades 3–8, 10) HWA (writing Assessment, Grades 4, 6, 9, 11)	Single scale score for Math and Reading along with normed percentile ranks	MC/OE Reading. and Math (grades 3–8, 10) + writing (grades 4, 6, 9, 11)	Released Items/ Practice Tests	Reading and Math, 4/9–4/20; Writing, 10/30–11/8.	Reading and Math, next fall = 16+ weeks; writing, mid-Feb.= approx. 13 weeks
Idaho	ISAT (Grades 3–8, 10)	Minimum of 12 items, Raw Scores + Scale Scores	MC	Practice Tests	4/16–5/11	Preliminary scale scores immediately, final score reports 1 week later
Illinois	ISAT 5 (Grades 3–8)	(Grade 3 Reading) Raw scores.	MC/OE + writing	Sample Items	3/12–3/23 Grades 3, 4, 5, 6, 7, and 8	Electronic 6/1 = 9 weeks
Indiana	ISTEP+ (Grades 3–10)	Scale Scores *(projected % based on projected underlying proficiency level—scaled score)	MC/OE + writing	Sample Items/ Released Items/ Released tests	09/19–9/30	1st week December (8–9 weeks)
	GQE (graduation exam)	(same as above)	MC/OE + writing	Sample Items/ Released Items/ Released Tests	09/20–9/22	1st week December (9–10 weeks)

Table 11.1: Fifty-State Reporting and Testing Information						
State	**Test**	**Reporting Categories + Minimum # Items for Strand Reporting—Raw Scores/Scaled Scores % Correct, If Applicable**	**Multiple-Choice/ Open-ended (+ Writing)**	**Sample Items, Released Tests, Practice Tests, Released Items**	**Test Dates/ Window**	**Score Reporting**
Iowa	ITBS (Grades 3–8) and ITED (Grades 9–12) and grades 3–8 have the option of taking the entire battery of ITBS or ITED. Tests are mandatory for Reading, Math, and Science for NCLB purposes. Optional Grades 3–8 tests include vocabulary, language, math computation, social studies, reference materials, and, maps and diagrams. At the high school, options include, vocabulary, spelling, revising written materials, social studies, and sources of information.	Minimum of 2 items = % correct; 3 or more items, also normative data	MC with locally determined options for CR supplements and writing.	NO	Early Sept to late April—school/ districts determine test dates	Paper and electronic results available within 13 days of receipt of materials

Table 11.1: Fifty-State Reporting and Testing Information						
State	**Test**	**Reporting Categories + Minimum # Items for Strand Reporting—Raw Scores/Scaled Scores % Correct, If Applicable**	**Multiple-Choice/ Open-ended (+ Writing)**	**Sample Items, Released Tests, Practice Tests, Released Items**	**Test Dates/ Window**	**Score Reporting**
Kansas	KMRA (Grades 3–8, HS)	4 items for substrands, reported as % correct.	MC (+ writing, scored locally—submitted and included in mid-May reports)	Released Items only for formative assessment	2/26–4/15	If tests taken online, immediately. If not, by May 15 = 4 weeks
Kentucky	CATS	No strand (subdomain) info on ISRs, but class rosters with correct/incorrect by item/strand; subdomain number correct reported by school	MC/OE + writing	Released Items/ Released Forms	4/16–4/27	mid-July = approx. 16 weeks
Louisiana	LEAP (Grades 4 & 8), *i*LEAP (Grades 3, 5, 6, 7), GEE (HS)	Minimum 4 items for *i*LEAP reporting	MC/OE + writing	Released Items/ Practice Tests	3/19–3/23	Electronic, 5/11; paper = 8 weeks

Table 11.1: Fifty-State Reporting and Testing Information						
State	**Test**	**Reporting Categories + Minimum # Items for Strand Reporting—Raw Scores/Scaled Scores % Correct, If Applicable**	**Multiple-Choice/ Open-ended (+ Writing)**	**Sample Items, Released Tests, Practice Tests, Released Items**	**Test Dates/ Window**	**Score Reporting**
Maine	MEA (Grades 3–8)	Report at cluster level for Math and text type for Reading. Percentage correct for school, district, and state. The minimum number of points reported is 5; the range is 5–28.	MC/OE + writing	Released Items/ Practice Tests	Grades 3–8, first 3 weeks of March; Grade 11, first 2 weeks of May	First week of July = 13 weeks (6 for Grade 11)
Maryland	MSA (Grades 3–8)	Overall Scale Score only for Math and Reading	MC/OE	Sample Items	03/12–03/21	approx. 7 weeks
	HSA (high school)	Overall Scale Score only for Math and Reading	MC/OE	Sample Items	05/21–05/25	approx. 9 weeks
Massachusetts	MCAS (Grades 3–9, 10)	Raw Scores by standard (+ scaled scores, grade 4 and up)	MC/OE + writing (Grades 4, 7, 10)	Released Items/ Released Tests/ Practice Tests	Reading 3/19–4/4; Writing, late March; Math/Science/Social Studies 5/14–5/31	Aug., preliminary reports(except H/SS), Sept., parent reports (except H/SS) =12 weeks +

Table 11.1: Fifty-State Reporting and Testing Information						
State	**Test**	**Reporting Categories + Minimum # Items for Strand Reporting—Raw Scores/Scaled Scores % Correct, If Applicable**	**Multiple-Choice/ Open-ended (+ Writing)**	**Sample Items, Released Tests, Practice Tests, Released Items**	**Test Dates/ Window**	**Score Reporting**
Michigan	MEAP (Grades 3–9) and MME (Grade 11)	ELA is reported in standard scores for overall ELA score (with accompanying performance levels)—scores are broken into a sub-scale score for Writing & Reading. Reading is broken into strands, reported as raw scores. Grade Level Content Expectations (GLCE) are reported as raw scores. Writing reported by strands as raw scores. Math is reported in an overall scale score with performance level descriptors. Raw scores are reported by sub-strand, and by GLCEs. Minimum for reporting 1 (e.g. spelling) at grade 3, most strands have significantly more items/points)	MEAP/MME MC/OE + writing	MEAP, Released Items; MME Practice tests	MEAP 10/8–10/26; MME mid-late March	Scores are made available in paper and available by PDF through a secure web site. A data research file is also available on the secure site. The CR writing student images is also available electronically on CD that is mailed to the district coordinator.

Table 11.1: Fifty-State Reporting and Testing Information						
State	**Test**	**Reporting Categories + Minimum # Items for Strand Reporting—Raw Scores/Scaled Scores % Correct, If Applicable**	**Multiple-Choice/ Open-ended (+ Writing)**	**Sample Items, Released Tests, Practice Tests, Released Items**	**Test Dates/ Window**	**Score Reporting**
Minnesota	MCA-II (Grades 3–8, 10)	Minimum 6–8 items, Raw Scores reported	MC/OE	Item Sampler	4/15–5/4	Normally, late August =14 weeks+
	Minnesota Writes!		Writing		Gr. 6 4/24–4/26	Reporting comes in with MCA-II
Mississippi	MSCT (Grades 3–8)	Minimum 4 items, Raw Scores reported	MC + separate writing test	Sample Items	5/1–5/3 (writing 3/6-–3/8)	June 15 = 5 weeks; writing July 13 = 13 weeks
Missouri	MAP (ELA Grades 3–8, 11; Math Grades 3–8, 10)	Minimum of 6 items, Raw Scores + percent	MC/OE + writing	Sample Items, Released Items	3/31–4/25	August 15 student scores arrive in districts = approx. 15 weeks
Montana	MontCAS (CRT-Reading/ Math Grades 3–8,10)	Minimum of 8 items, Raw Scores + percent	MC/OE (math)	Sample Items, Released Items, Released Tests	3/5–3/28	"Before the end of the school year"
Nebraska	STARS (School-led Teacher-based Assessment and Reporting System). No state-developed test, except for writing.	N/A	N/A	N/A	Districts report local assessment results to the state by June 30th	State results reported in fall. Districts report locally to parents/educators—no predetermined schedule.

Table 11.1: Fifty-State Reporting and Testing Information						
State	**Test**	**Reporting Categories + Minimum # Items for Strand Reporting—Raw Scores/Scaled Scores % Correct, If Applicable**	**Multiple-Choice/ Open-ended (+ Writing)**	**Sample Items, Released Tests, Practice Tests, Released Items**	**Test Dates/ Window**	**Score Reporting**
Nevada	CRT (Grades 3–8); HS Prof. Exam	Raw Scores and Percent for content strands and cognitive levels. Minimum of 6 items for reporting.	MC/OE (Grades 4–8 only) + writing	Practice Tests	Grades 3–8 approx. 3/1–4/13; HSPE 3/26–3/30; Writing Grade 5 1/16–1/26, Gr. 8 2/5–2/15	4 weeks electronic/ approx. 6 weeks paper. Writing on a similar schedule.
New Hampshire	NECAP (Grades 3–8)	Raw Scores for subcategories. Minimum of 10 points.	MC/OE + writing	Released Items/ Practice Tests	10/2–10-24	Electronic reports available 1/29 = approx. 12 weeks; paper shipped week of 2/5 = approx. 13 weeks
New Jersey	NJ ASK (Grades 3–7), GEPA (Grade 8), HSPA (High School)	Raw Scores for clusters. Minimum of 7 points.	MC/OE + writing	Released Items/ Sample Tests	Various windows between 3/05–3/30	mid-June = approx. 10 weeks
New Mexico	NMSBA (Grades 3–9, 11)	The substrand results are provided in hard copy reports—Raw Scores and % Correct Desired minimum is 6 (DOE working to correct some being reported with as few as 3.)	MC/OE + writing	Released Items	2/26–3/23	8/1 = approx. 17weeks

Table 11.1: Fifty-State Reporting and Testing Information						
State	**Test**	**Reporting Categories + Minimum # Items for Strand Reporting—Raw Scores/Scaled Scores % Correct, If Applicable**	**Multiple-Choice/ Open-ended (+ Writing)**	**Sample Items, Released Tests, Practice Tests, Released Items**	**Test Dates/ Window**	**Score Reporting**
New York	ELSA/ILSA (Grades 3–8) [HS Regents Exam info not included]	Yes, but reported as a Standard Performance Index—projects how a student would do if there were 100 items in a standard, based on item difficulty. Minimum of 3 points used.	MC/OE and ER	Sample Tests	ELA – January; Math – March	ELA – May 12 + weeks; Math – June 8+ weeks
North Carolina	EOG (Grades 3–8)	Goal reporting with 8 items minimum, Raw Score and mean % correct reported.	MC + writing	Sample Items	Last three weeks for EOG (up to 7-day testing window). Writing, 2nd Tuesday in March.	Almost immediately at district-level for MC, as tests are locally scanned and scored—paper reports approx. 1 week later. Approx. 8 weeks for writing scores from contractor.

Table 11.1: Fifty-State Reporting and Testing Information						
State	**Test**	**Reporting Categories + Minimum # Items for Strand Reporting—Raw Scores/Scaled Scores % Correct, If Applicable**	**Multiple-Choice/ Open-ended (+ Writing)**	**Sample Items, Released Tests, Practice Tests, Released Items**	**Test Dates/ Window**	**Score Reporting**
North Dakota	NSDA (Reading/ELA and Math, Grades 3–8 and 11)	Reports out the number of items and points for a standard (and for a benchmark), and the percentage of points earned by at each standard and benchmark. Reporting for each benchmark, even if there is only one item for that benchmark, with cautions to schools not to make a judgment related to that benchmark, but to consider it as part of the particular standard.	MC/OE	None	10/22–11/9	Last week February = 14+ weeks
Ohio	OAT (Grades 3–8) and OGT (HS grad exam)	9-point minimum for reporting at strand level, reported at one of three performance levels.	MC + writing grades 4 and 7	Practice Tests/ Released Items	4/28–5/9 (Grades 3–8); 3/10–3/23 (HS spring grad exam)	All within 60 days (8 weeks) of testing.

Table 11.1: Fifty-State Reporting and Testing Information						
State	**Test**	**Reporting Categories + Minimum # Items for Strand Reporting—Raw Scores/Scaled Scores % Correct, If Applicable**	**Multiple-Choice/ Open-ended (+ Writing)**	**Sample Items, Released Tests, Practice Tests, Released Items**	**Test Dates/ Window**	**Score Reporting**
Oklahoma	OCCT (Grades 3–8); OCCT End of Instruction (HS)	Student level reporting includes percent correct at the standard and objective level, with no minimum number of items/points. Summary reporting includes the percent correct at the standard and objective level, with a minimum of 4 items (4 points) required for reporting.	MC + writing grades 5 and 8, English II	Practice Tests (Grades 3–8); Sample Items (EOI)	Grades 3–8, MC only, 4/10–4/27 for both p/p and CBT. EOI piloting, 5/7–5/18. Once operational, the EOI tests will have 4 separate windows each year.	Paper/pencil: Preliminary reports by 6/1 = approx. 5 weeks; Final reports by 7/1 = approx. 9 weeks CBT: Student raw scores, immediately upon test completion; Student-level results, within two weeks of the end of the testing window (legislatively mandated by 2008–2009) Preliminary and Final reports same as for p/p.
Oregon	TESA (Grades 3–12)	YES—Depending on whether a long or short form of the test is taken, reporting of strands may be provided at the student level. Minimum of 8–10 items.	MC	Released Items/ Sample Tests	September–May	Immediately—tests are online; system-level reports in August

Table 11.1: Fifty-State Reporting and Testing Information						
State	**Test**	**Reporting Categories + Minimum # Items for Strand Reporting—Raw Scores/Scaled Scores % Correct, If Applicable**	**Multiple-Choice/ Open-ended (+ Writing)**	**Sample Items, Released Tests, Practice Tests, Released Items**	**Test Dates/ Window**	**Score Reporting**
Pennsylvania	PSSA (Grades 3–8, 11 Reading/Math; Writing Grades 5, 8, 11)	Raw Scores reported at the strand level: minimum of 8 items.	MC/OE + writing	Sample Items/ Released Items	2/12–2/23 Writing, Grades 5, 8, 11; 3/12–3/23 (Grades 3–8, 11 Reading/Math)	Electronic reports to districts, 7/9 = approx. 18 weeks for writing, approx. 14 weeks for MC; hard copies 9/1
Rhode Island	NECAP (Grades 3–8)	Raw Scores for subcategories. Minimum of 10 points.	MC/OE + writing	Released Items/ Practice Tests	10/3–10/25	Electronic reports available 1/29 = approx. 12 weeks; paper shipped week of 2/5 = approx. 13 weeks
South Carolina	PACT (Grades 3–8)	Reporting at Content Area – Total scale scores and performance levels for each of the four subjects. For ELA, students also receive a performance level for the reading and writing.	MC/OE + writing for ELA, MC/OE for Math, MC only for Science and Social Studies	Released Items by subject	ELA – 5/7and 5/8 Math – 5/9 Science – 5/15 Social Studies – 5/16 Field Testing – 5/17 Make-up testing through 5/22	Preliminary Below Basic Reports—staggered by grade from 6/18–7/16 Electronic reports staggered by grade from 6/22–7/31 with final CD-ROM on 7/31 Hard copy reports on 8/23

Table 11.1: Fifty-State Reporting and Testing Information						
State	**Test**	**Reporting Categories + Minimum # Items for Strand Reporting—Raw Scores/Scaled Scores % Correct, If Applicable**	**Multiple-Choice/ Open-ended (+ Writing)**	**Sample Items, Released Tests, Practice Tests, Released Items**	**Test Dates/ Window**	**Score Reporting**
South Carolina (continued)	HSAP (High School)	Reports scale scores and at achievement levels (1–4) by content area	MC/CR and ER for ELA, MC/CR for Math	Prototype forms for both content areas ELA release items, answer key, and rubrics Math release items and answer key	ELA – 4/24 and 4/25 Math – 4/26 Make-up testing through 5/4 ELA and Math subtests – 7/24–7/26	Electronic and print score reports are delivered to districts approx. 2 months after testing with the exception of expedited reporting for graduating seniors. Spring – 5/15 and 7/13 Summer – 8/17 Fall – 12/15
	EOCEP	Reports scale scores and letter grades for each subject	MC only	Sample items and keys are included in Teacher's Guide	Fall – 12/1/06–1/31/07 Spring – 5/1/07–6/7/07 Summer – 6/18/07–8/3/07	Student results 36 hours after contractor receives electronic or paper responses; school/district/ state summaries, 4–5 weeks
South Dakota	Dakota STEP (Grades 3–8, 11)	The standards and indicator level-based reports provide percentages, scaled score, and raw score. Minimum number of items is 7.	MC	No	4/2–4/20	Available to schools August = 13 + weeks
Tennessee	TCAP (Grades K–8)	Reported by category with a Performance Level and a Performance Index Score. Minimum number of items for reporting unknown.	MC + writing (Grades 5 and 8)	Item Sampler	4/2–4/25; writing 2/6	72 hours electronic/ 8 weeks paper

Table 11.1: Fifty-State Reporting and Testing Information						
State	**Test**	**Reporting Categories + Minimum # Items for Strand Reporting—Raw Scores/Scaled Scores % Correct, If Applicable**	**Multiple-Choice/ Open-ended (+ Writing)**	**Sample Items, Released Tests, Practice Tests, Released Items**	**Test Dates/ Window**	**Score Reporting**
Texas	TAKS Grades 3–11 (for AYP grades 3–8 and 10) Reading, math, science, writing/ English language arts, social studies (for AYP reading, math, science)	Reporting at Objective level, minimum of 4 items per objective, Raw Score	MC + writing and OE for HS ELA	Released tests (all grades and subjects, bi-annually), plus Sample Items.	2/20–2/22 (Writing 4, 7 = Multiple Choice + composition Reading 9 = Multiple Choice + Open-ended English language arts = Multiple Choice + Open-ended + composition 4/16–4/20 (MC-only tests)	5/4–5/11 = 2–3 weeks for MC only, approx. 12 weeks writing/OE. Note: High stakes tests (for students) are reported ten working days after the testing contractor receives the scorable materials (Grade 3 reading, grade 5 reading and math, exit level math, ELA, science, social studies)
Utah	CRT (Grades 3–8)	Raw scores by subcategory (ILOs). Minimum of 4 items.	MC	No	The last 5 weeks before the end of the school year.	Raw score reports are provided to districts four days after the district submits student answer documents to the USOE. (Online tests give raw scores in 48 hours)
Vermont	NECAP (Grades 3–8)	Reporting by "subcategory" in Math, Reading, and Writing. Reported by raw scores with all subcategories having a minimum of 10 raw score points	MC/OE+ writing	Released Items/ Practice Tests	First three weeks of October	electronic reports available 1/29 = approx. 12 weeks; paper shipped week of 2/5 = approx. 13 weeks

Table 11.1: Fifty-State Reporting and Testing Information						
State	**Test**	**Reporting Categories + Minimum # Items for Strand Reporting—Raw Scores/Scaled Scores % Correct, If Applicable**	**Multiple-Choice/ Open-ended (+ Writing)**	**Sample Items, Released Tests, Practice Tests, Released Items**	**Test Dates/ Window**	**Score Reporting**
Virginia	SOL (Grades 3–8. and HS EOCs)	7 item minimum, Raw Scores reported—VA provides data at the subject level and at the reporting category level (strand)—raw scores and scaled scores. We provide a "Student Performance by Question Report" that indicates how the student performed on each item and a brief descriptive phrase about the content of the item. The SPBQ Report, as it is called, is also summarized at the school, division, and state levels at the end of an adminis-tration so a school division can see at the item level how a student, school, their division, or the state performed on individual items.	MC + online writing assessment planned for Grade 8	Released Tests (only released items for History tests)	Spring Non-Writing: Divisions [districts] may set their 2-3 week EOC testing window within the statewide testing window. Statewide Spring window is 4/16–6/29. Divisions must select one of three possible windows for their 3–8 tests administered in the spring. Options for Grades 3–8 windows are: 4/16/07–5/11/07; 5/7/07–6/1/07; and 5/21/07–6/15/07.	For a previously post-equated form OR once a new test form has been post-equated, a student's results for online tests are available immediately after the student completes the online test. The immediate results for online tests are in the form of a downloadable and printable page with the total and reporting category raw and scaled scores. By the following morning, a data file with all demographic data, all raw scores, and all scaled scores (total and reporting category) is available. For unequated forms, divisions [districts] wait for post-equating to occur, usually no longer than 4 weeks for p/p and 2 weeks for online tests.

Table 11.1: Fifty-State Reporting and Testing Information						
State	**Test**	**Reporting Categories + Minimum # Items for Strand Reporting—Raw Scores/Scaled Scores % Correct, If Applicable**	**Multiple-Choice/ Open-ended (+ Writing)**	**Sample Items, Released Tests, Practice Tests, Released Items**	**Test Dates/ Window**	**Score Reporting**
Washington	WASL (Grades 3–8)	Scale scores/ performance levels by subject	MC/OE + writing (HS only)	Released Items, Sample/Practice Tests	Grades 3–8, 4/16–5/4; HS (Reading/Writing.) 3/13–3/20 (Math Science) 4/17–4/20	All scores released early Sept. = approx. 16 week; HS by 6/14 to parents = approx. 7 weeks
West Virginia	WESTEST (Grades 3–8, 10)	Individual Item analysis summary provided. % of students who answered item correctly, mean %, and # of points possible.	MC/OE + writing at 4, 7, 10 (the latter online for 7 and 10)	Released Items	MC/OE 5/14–5/18; writing, varies, late February to late March	8 weeks for all tests
Wisconsin	WKCE-CRT (Grades 3–8, 10)	Minimum of 5 items for a reporting category. Reported as a Standard Performance Index—projects how a student would do if there were 100 items in a standard, based on item difficulty.	MC/OE + writing	Released Items/ Sample Items/ Practice Tests	10/23–11/24	10–12 weeks paper

Table 11.1: Fifty-State Reporting and Testing Information						
State	**Test**	**Reporting Categories + Minimum # Items for Strand Reporting—Raw Scores/Scaled Scores % Correct, If Applicable**	**Multiple-Choice/ Open-ended (+ Writing)**	**Sample Items, Released Tests, Practice Tests, Released Items**	**Test Dates/ Window**	**Score Reporting**
Wyoming	PAWS (Grades 3–8,11)	Substrand reporting (8 for Reading, 12 for Math), with an absolute minimum of 5 items and preference for 7–8. Raw score and Scale score.	MC/OE + Writing	Released items with data	Five week testing windows (Spring window = p/p 3/26–4/19; CBT 3/26–4/27)	During the testing window for MC. Four weeks after the close of the testing window for points of points possible MC an OE. Eight weeks for proficiency and scale scores.

COMPONENT 11 – REFERENCES AND LINKS

American Institutes of Research, Ohio Reports. Available at http://ohiokids.air.org/

Trout, D., and Hyde, E., (2006) "Developing Score Reports for Statewide Assessments that are Valued and Used: Feedback from K–12 Stakeholders," Pearson Educational Measurement. Available at http://www.pearsonedmeasurement.com/research/c_papers.htm#

Council of Chief State School Officers (CCSSO)'s Formative Assessments in a Comprehensive Assessment System Study Group; summary available only at http://www.ccsso.org/Projects/scass/projects/comprehensive_assessment_systems_for_esea_title_i/study_groups/1960.cfm

CTB/McGraw Hill and the Grow Network. Available at:

http://www.mcgraw-hill.com/releases/education/20070315.shtml

http://www.ade.state.az.us/asd/statetutoring/aimshstutoringguide.asp

http://growparent.com/

http://www.yourstudyguide.com/download/init.do?gjid=22290580-17046615651179502674182&market=TXS

Luecht, R.; Gierl, M.; Tan, X.; & Huff, K. (2006, April). *Scalability and the Development of Useful Diagnostic Scales.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA.

COMPONENT 12 – ADVANTAGES AND DISADVANTAGES OF CONVERTING THE STATE ASSESSMENT PROGRAM TO A COMPUTER-BASED OR COMPUTER-ADAPTIVE ASSESSMENT

This component of the study requires an analysis of the advantages and disadvantages of computerized testing, focusing on both CAT and CBT. DRC proposed that one important source for obtaining the latest, most accurate information about states' initiatives in computerized testing would be to convene a panel of experts and researchers representing states that have implemented CAT or CBT. The panel would discuss the advantages and disadvantages of moving to CAT/CBT assessments and offer lessons learned from their state initiatives.

DRC asked the SC Team to recommend states that should be invited to serve on this panel. The SC Team indicated a particular interest in eliciting advice from southeastern states that had implemented some type of CAT/CBT, as well as other states that had used a variety of approaches to roll out their computerized testing plans. To ensure the expert panel's assessment programs represented a wide array of implementation strategies and CAT/CBT vendors, DRC used results from our Internet research and survey of state computerized testing initiatives to identify thirteen states as potential expert panel participants. The states invited were Mississippi, Idaho, Florida, North Carolina, Oklahoma, Virginia, Wyoming, Texas, Oregon, Georgia, West Virginia, Indiana, and Kentucky. Each of these states has approached computerized testing in a different manner to attain varied goals. Additionally, the managing editor of *Technology Horizons in Education Journal* was invited; however, he was unable to attend.

Of those invited, the following states sent representatives to the Expert Panel meeting held on March 28, 2007, in Columbia, SC:

- Florida—Kristen Ellington
- Idaho—Karen Echeverria
- Indiana—Wes Bruce
- Kentucky—Roger Ervin
- North Carolina—Mildred Bazemore
- Oklahoma—Jennifer Stegman
- Virginia—Sarah Susbury
- West Virginia—Sandra Foster

Additionally, the three feasibility study consultants—representing the University of Kansas, James Madison University, and the University of North Carolina-Greensboro—participated in the meeting. The meeting was facilitated by Dr. Steven Wise, a consultant to this project and a Senior Assessment Specialist at James Madison University's Center for Assessment and Research Studies.

Other meeting attendees included the state of South Carolina team members, members of the Advisory Committee, and DRC staff whose role was to manage meeting logistics and serve as recorders. The meeting was posted as an open meeting and this posting met the 24-hour-prior-to-the-meeting deadline, as required by South Carolina statute.

Tami Mainwaring, Project Leader of The Team (Change Manager, Division of the State CIO) welcomed the group and discussed the background for, and purpose of, the feasibility study. Patricia Porter, Vice President of Large-Scale Assessment for DRC, gave a brief overview of the fourteen components delineated in the Request for Proposals. Dr. Wise reviewed the agenda and asked panel members to provide a brief description of their computerized testing initiatives, to reflect upon lessons learned from their own experiences, and to offer advice to South Carolina.

Topics for discussion included the following:

- Overview of each represented state's computerized testing program
- Implementation plans
- Computer adaptive vs. fixed/scrambled forms
- Test design and item types
- Accommodations
- Use of handheld devices/emerging technologies
- Testing windows
- Security
- Staffing and training
- Funding sources
- Costs to state and districts
- Management and support of student databases

- Availability of computers and impacts on instruction
- Risk planning

Advice offered by Expert Panel members included the following points, to which panel members expressed general agreement:

- Set the expectations and the goals to be achieved through CBT or CAT.
- Ensure that districts and schools have the necessary technology, bandwidth, hardware, and training to be successful in implementing the program.
- Make certain that the Request for Proposal is very specific and details the minimum threshold of PC requirements.
- Lay out an implementation plan that has incorporated input from all stakeholders.
- Make certain that the legislature is prepared to fund the up-front costs.
- Select a knowledgeable committee to serve as ongoing advisors for the computerized testing initiative.
- Call for an independent third-party assessment of the chosen vendor's security and system.
- Build the assessment specifically for online delivery.
- Use scientifically-based research and psychometric "best practices" to guide decisions.
- Plan for sufficient technology support.

- Ensure that training for districts is a team effort between the state and the vendor.
- Prepare for some resistance.
- Elicit advice from other states with computerized testing programs.
- Start with a practice or pilot test before implementing high-stakes computerized assessments.
- Link formative or practice tests to the online summative assessments.
- Inform policymakers and stakeholders that initial costs will probably be higher than with a completely paper-based assessment system.
- Foster the idea that the way students are tested should always reflect the way they are taught (i.e., the use of technology should be an integral part of classroom instruction).

Official minutes from the meeting were taken by Ms. Mainwaring. The minutes of this meeting are included in the Appendices section (see Table of Contents).

The section below describes advantages and disadvantages of converting the South Carolina state assessment program to a computerized system, whether CBT or CAT. Information was garnered from the Expert Panel, questionnaires to all states, literature reviews, and the collective experiences of computerized-testing researchers and practitioners.

Computer-Based Tests/Testing: Advantages and Disadvantages

IMMEDIACY OF RESULTS

Among the most notable features associated with computerized testing is the potential for immediate test results. To obtain immediate results, however, requires that the assessment is either comprised entirely of multiple-choice or multiple-choice with short-answer test items that can be scored objectively using artificial intelligence systems. In such a situation, as soon as a student finishes taking a test, the scores could be immediately available for reporting. Immediate availability of results may be considered counterproductive if districts and/or local media begin compiling their own interpretations of test results before the state department of education can officially release comments and/or any statewide results to the media.

A somewhat related question is to whom the immediate results should go. Should they be provided to the student at the conclusion of the test? Should the results be provided only to the teacher, who can then review them with the student? What type of results should be immediately reported?

As described in Component 11, many states have chosen not to abandon the assessment of written composition in order to expedite the reporting of results, but instead administer their writing assessments earlier in the school year, through either conducting a separate writing test or administering their entire English language arts test earlier. These adjustments to the typical end-of-year assessment provide sufficient time for the compositions to be scored by trained readers. So far, only one state, West Virginia, has opted to have student compositions scored using artificial intelligence.

Providing immediate test results also typically requires that the psychometric characteristics of the test items be based on pilot test/field test data. This is referred to as "pre-equating" and has some potential drawbacks, particularly if the field-testing is done at a separate time from the operational test, since there are often concerns that students may not have put forth optimal effort when taking a separate field test. Since field-testing typically involves the piloting of many items and/or tries to minimize student time away from instruction, the psychometric characteristics of a field test are typically based on a relatively small number of students. Although the number of students participating in field-testing is typically quantified so as to yield useful item-level data, and the field-test sample is drawn to be representative of state demographics, many states still engage in updating the psychometric information gleaned from field-test items after operational testing has occurred. Thus, items statistics will be gathered on a much larger sample taken from students testing under optimal motivation conditions (i.e., "live" testing). This is referred to as "post-equating" and many states, including South Carolina, use a robust post-equating model in their statewide testing programs.

Continued use of a post-equating model does not preclude the possibility of computerized testing, but it does preclude the possibility of immediate results. Results might still be faster than what could be achieved under standard paper-based testing, but the results would not be available immediately following testing without a shift to a pre-equated model. The state would need to carefully consider this shift in paradigms from post-equated to pre-equated models and discuss this with its Technical Advisory Committee to evaluate how this paradigm shift would impact the transition from paper-based to computerized testing and/or the co-existence of both paper-based and computerized tests in the state.

If constructed-response or extended-response items, such as writing a composition, remain in the South Carolina assessment program, and experienced human scoring continues to be utilized, the state might consider implementing computerized testing with minimally delayed reporting. That is, using computers to administer the entire test, the constructed-response and/or extended-response items would be entered via computer and then routed electronically to a scoring center where they would be scored by human readers. Such a system would ensure the careful application of the scoring rubric, reader/scorer training, and the educational requirements of readers/scorers that the state department of education requires.

As soon as a student has all components completed through the scoring process, the combined total test score would then be available for reporting. This availability also depends on a pre-equated psychometric model. It does, however, have the potential for transitioning to computerized testing while maintaining the current test design. This set of procedures would capitalize on the logistical advantages of computerized testing as well as reduce turnaround time for score reporting.

COST EFFICIENCY

Moving the South Carolina assessment program from a system that is paper-based to one that is delivered almost exclusively via computer will likely not result in a significant reduction in overall assessment costs, particularly in the early years of the computerized assessment program. Even though the production of test booklets and answer documents and the shipment of large quantities of materials will be greatly reduced—and eventually almost eliminated—certain components of any high-stakes assessment program must remain, irrespective of the mode of delivery. Specifically: development of test specifications; item development and review; field-testing; test security procedures; test form development; human scoring of CRs/ERs (should these be retained); psychometric analyses to ensure proper equating and standard-setting; and creation of paper score reports for distribution to districts, schools, and parents.

Additionally, a transition to computerized testing would require comparability studies, as required by NCLB. Finally, the initial investments in hardware, software, capacity, and staffing will obscure cost savings in the early years of computerized test administration, but may be offset by enhancing access to technology in the classroom and allowing technology to become a more integral part of classroom instruction.

Instructional Uses

Once schools have the necessary expertise and capacity for computerized testing, the infrastructure (hardware, software, connectivity, networking, and technology expertise in all districts) can be utilized on a daily basis to support and extend instruction. Students will have access to Internet research, be able to perform online science experiments, compose essays on the computer, and take self-paced online courses. Teachers will be able to administer formative assessments and use the results to inform their instructional practices, monitor their students learning progress through an online database, and create innovative lesson plans that maximize the use of technology.

Usefulness and Accuracy of Student Data

The speed with which data are available to educators allows rapid analysis and evaluation of performance. Whereas in the past it could take three months for data to be returned to administrators and teachers, today's computerized tests allow educators to review results and make instructional and evaluative decisions on behalf of students while the information is current. When students take a test on a computer, the storage and retrieval of information is prompt and efficient. In the past, reliance on answer sheets meant data checking, scanning and processing, scoring and reports preparation, and eventual readying of information for storage and retrieval.

Computerized testing allows the collection of additional information, (e.g., student surveys, the time taken to answer each item). Student results from an online assessment system can be easily imported into a state- or district-supported data warehouse. Schools could easily disaggregate their data in meaningful ways, tie it to attendance rates and other pertinent demographic variables, calculate Adequate Yearly Progress ratings, and provide teachers with up-to-the-minute student data from formative and benchmark assessments. Additionally, student-level information captured online prior to the assessment will likely minimize student errors in miscoding information on test answer documents.

STUDENT MOTIVATION

An additional advantage of computer-based testing lies in its potential for addressing issues concerning student test-taking motivation. There has been a longstanding concern that some students may not be motivated to perform on tests for which there are no individual consequences (e.g., course grade, high school diploma, college entrance). The validity of an inference made about a student based on a test score, however, is dependent on the degree to which the student gave his or her maximal effort. If the student is not motivated, test performance is likely to suffer, resulting in a test score that underestimates the student's true level of proficiency.

Computer-based tests can measure the amount of time students spend on each test item, which Wise and Kong (2005) showed can be used to construct a valid measure of student test-taking effort. In a similar fashion, Wise (2006) showed that response time can be used to measure the amount of effort received by different test items. Both of these measures could be used to improve test score validity by identifying the circumstances under which students are not motivated, and the types of test items that appear to elicit the greatest amount of effort from students. Additionally, computers could be used to monitor student effort as a test is being administered, displaying messages of encouragement or warning to those students exhibiting low effort. Wise, Bhola, and Yang (2006) studied this type of effort-monitoring test and found that it yielded test scores with higher validity than those from a traditional computer-based test.

Computer Adaptive Component to Support-Targeted Instruction

Moving traditional tests to computerized administration has opened the door to CAT, which is especially useful in targeting academic strengths and weaknesses of both above-grade-level and below-grade-level students. However, as explained in Component 3, a CAT system will not meet the requirements of NCLB unless all the items used for federal reporting purposes are on-grade-level and aligned to state standards. In spite of this restriction, a traditional CAT system can be useful as a formative assessment to target specific content standards/strands in classroom instruction and to informally measure students' learning.

Since a CAT used as a formative instrument could potentially reduce the length of traditional tests by 50–75%, it would likely be viewed as an aid to instruction rather than an impediment. Additionally, even though CAT would permit considerable reduction in test length, there is a relatively small loss in terms of validity and reliability.

Construction of a CAT in which the items in the bank span a range of grade levels requires that the items be vertically scaled. There are some differences of opinion in the psychometric community about the utility of vertical scales in achievement level testing. For example, there are fundamental differences in life and academic experiences between a 10th grade student whose achievement is at a 3rd grade level and an actual 3rd grade student. The vertical scale assumes that both students would approach an item with the same instructional background and skill sets. Clearly, this is an area where expert opinion may differ, and the state is advised to gather advice from multiple sources and to proceed cautiously in making decisions regarding the implementation of a CAT that uses off-grade items.

Reduced Administrator and Instructor Effort

As stated above, if South Carolina moves to computerized testing, the traditional activities of receiving, unpacking, securing, counting, sorting, and distributing test booklets and answer sheets that are performed by administrators and teachers would be virtually eliminated. However, the management functions of scheduling, monitoring, and implementing the computerized testing sessions will still be needed.

Training will be needed on how to use the computerized testing system, how to get students started to take the computerized tests, and how to handle common problems. This may simply represent a shift in the types of preparation that administrators and teachers need to have. If the testing will take place in computer lab settings, training will also need to focus on someone other than the teacher serving as the test administrator. In some situations, the credentials of individuals serving as test administrators need to be considered, as some states require that the test administrator be a certified/licensed teacher.

MEETING THE NEEDS OF SPECIAL POPULATIONS

Technological advances in computerized testing signal a new era in testing for students with special learning needs and other populations, such as English Language Learners. Administering accommodated tests (e.g., oral administrations) via paper/pencil is typically labor-intensive and usually necessitates additional test administrators and testing rooms. In many instances, accommodations to paper-pencil tests have been implemented in an unstandardized manner. A research-based computerized testing system can potentially offer an array of accommodations, such as increasing text size with the click of a button, transforming text to speech, zooming, colorizing objects, and providing students or test administrators with tools to change the display on the screen for sections of text and test items, when warranted. These accommodations could also be accomplished automatically by collecting information regarding needed accommodations prior to testing, possibly in conjunction with the administration of the practice test.

ABILITY TO MODIFY TESTS

In the past, when an error was discovered within a printed test booklet, there may not have been sufficient time to correct the error, reprint testing materials, and get them where they needed to be for testing. Within the context of a computerized testing system, errors can be corrected much more quickly since the logistical aspects of printing, shipping, and distribution within the district are not an issue. There also exists the possibility of correction of errors found during the course of actual test administration. This latter scenario, however, would need to be carefully considered. One could correct the error to minimize impact of the error on other students but, in that case, there is still the issue of how to address those students who have already tested using a booklet containing the error. Such decisions are difficult and multi-faceted, and must be addressed by the South Carolina Department of Education.

IMPROVED TEST SECURITY

As tests are delivered via computer, traditional handling of test materials is eventually all but eliminated; thus, South Carolina breaches of test security may be fewer. However, other types of test security issues, such as the potential of hacking into the system, need to be addressed. To date, there have generally been few reports of security issues due to hacking. Vendor and district/school security procedures are paramount in avoiding such breaches.

One additional new area of potential test security breaches with computerized testing is that a longer window of time will be needed for the test administration because not every student will have simultaneous access to a computer to take the test. This situation creates potential issues with students sharing information, perhaps innocently, about the test with other students who have not yet tested. This potential issue can be addressed with either multiple parallel test forms or CAT. Another possible area to consider is reordering the sequence of test questions within a test to avoid having all students begin the test and see the same initial question appear on the monitor.

BARRIERS TO IMPLEMENTATION

Local resistance to change—Change can come very slowly for some individuals. The use of new technology brings with it a necessary requirement to master new skills. In addition, some individuals may see no need to change the way testing has previously been done. Careful, thoughtful steps must be planned to assist individuals making the transition. Without assistance, the acceptance of change may be very challenging for a few individuals. These persons would need to be made to feel comfortable and supported as the transition to computerized testing proceeded. Additionally, to the extent possible, school district staff, such as school and district technology coordinators, should visibly and actively support the computerized testing initiatives.

Local capacity—A school or school district may not have the technological capacity to move to computerized testing. Hardware and software may be outdated or lacking altogether, and the local staff may not have the technological expertise to implement and monitor a successful computerized testing program.

Mitigation of risks and obstacle to success—Due to the high-stakes nature of the tests, failure and errors should not be tolerated, and plans must be prepared in advance to deal with unforeseen situations.

Perceived impediments to implementation of computerized testing will add to frustration and dissatisfaction. When systems are new to the teacher and testing is pressing forward within the scenario of high-stakes accountability, tolerance for the unexpected will be very low. Make the systems understandable and forgiving.

Four Keys to Success

1. **Communication:** Frequent, brief, and effective communication is essential. Share information about the computerized testing program and its activities, benefits, and cautions. Keep all stakeholders well informed, build a community of knowledgeable people, and ensure that educators are involved in all key steps. Also, ensure that adequate communication is shared with all stakeholders about the successes and failures encountered as the transition to computerized testing proceeds.
2. **Involve technological staff:** District and school level technology staff needs to become involved from the outset of any transition to computerized testing. Provide assistance to local technology staff on how to assist district and school staff and educators. The local technology staff is key to ensuring the success of the computerized testing program.
3. **Training workshops:** Plan training sessions to be conducted periodically during the semester. Get teachers working at the same grade and across grades solving problems together. These training sessions can be brief (perhaps limited to one or two hours each).

4. **Training materials:** The state should seriously consider the development of student and teacher tutorials regarding use of the computerized testing system, student practice tests, and implementation guides for teachers and school/district administration staff.

COMPONENT 12 – REFERENCES AND LINKS

Wise, S. L. (2006). An investigation of the differential effort received by items on a low-stakes, computer-based test. *Applied Measurement in Education, 19,* 95–114.

Wise, S. L., Bhola, D., & Yang, S. (2006). Taking the time to improve the validity of low-stakes tests: The effort-monitoring CBT. *Educational Measurement: Issues and Practice 25(2),* 21–30.

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education, 18,* 163–183.

COMPONENTS 13 & 14 – A REASONABLE IMPLEMENTATION SCHEDULE AND CONCLUSIONS AND RECOMMENDATIONS

The recommendations presented below regarding the possible implementation of computerized testing in South Carolina schools target the five key goals that have been articulated by the state's policymakers and educators. South Carolina's goals are similar to those expressed by other states that have or are considering the implementation of computerized testing.

The goals are the following:

- the desire to receive assessment results more quickly;
- the need to address concerns about the amount of time spent on testing;
- the need to obtain diagnostic information from assessment results;
- the need to fully understand the costs associated with computerized testing; and
- the desire to provide an instructionally, psychometrically sound, and useful assessment system of the highest quality.

These goals support the need for developing a computerized assessment system that is based on solid research, will meet federal and state requirements, provides sufficient information to educators to guide them in targeting student instruction, and is fiscally prudent.

The implementation of computerized testing in the state can be a significant factor in helping South Carolina to meet its key goals. This study reports the current computer capacity of South Carolina and projects that teacher/student use of computers in the classroom will increase markedly over the next ten years. As computers become increasingly integral to instruction, it is appropriate that they become correspondingly integral to assessment. The link between instruction and assessment is paramount.

There are a number of advantages, as well as barriers, associated with electronic testing, which are discussed in Component 12 of this report. Delineated below are recommendations that are grounded in the findings presented throughout this report, along with a reasonable schedule for implementation. Data sources included the following:

- Surveys of all 50 states
- Surveys of readiness of South Carolina districts and schools
- Literature reviews of score comparability when different assessment modes are used
- A review of pertinent NCLB requirements and their application to computerized testing
- Advice from the Expert Panel
- Research on the efficacy of including constructed response items
- Implementation costs gleaned from other state programs and analyses of necessary computer infrastructure
- Pertinent South Carolina legislation

Recommendations for Implementation

Implement computerized testing in South Carolina using a multi-year/multi-phase rollout plan.

As stated throughout this report, more and more states are moving to computerized testing, and the medium for administering statewide assessments is evolving from a paper-based system to one that is delivered electronically. Since electronic media are being utilized more and more in classroom instruction, the mode of assessment must reflect the mode of instruction, as was the case in the paper/pencil days. However, as described in Components 7 & 9, it is critical that school districts have the capacity, such as necessary hardware, software, and infrastructure, as well as sufficient knowledgeable staff, to ensure a smooth transition to computerized testing. In order for districts to be able to reach this capacity, they must be provided with necessary funding and sufficient time to ensure adequate resources prior to full implementation. A recommended implementation plan, projected costs, and assumptions upon which the costs are based are described below. *This is a reasonably aggressive plan that may need to be implemented at a slower pace to align with available monetary and staff resources and to accommodate early implementation successes or challenges, evolving state and federal requirements, and state/district/school support for computerized testing.*

Phase One

WHAT SUBJECT AREA?

Science

Rationale: As with any new initiative, starting off successfully goes a long way toward ensuring long-term success. Therefore, it is recommended to begin with a single subject area and that the initial subject area for statewide implementation should be one that is not currently used in NCLB Adequate Yearly Progress (AYP) calculations. Science assessment results are not included in these calculations at this time. It is likely that South Carolina school districts would be more likely to want to participate in a computerized science assessment, since the stakes for schools and districts are not currently quite as high for science as they are for mathematics and English language arts in terms of NCLB reporting.

Please note that recent information disseminated by the U. S. Department of Education Secretary Margaret Spellings in her April 23, 2007, letter to Chief State School Officers calls for performance data on science assessments to be included in AYP accountability calculations in the future. One must remember, however, that science assessment results are used currently in South Carolina's own accountability system. Thus, the non-NCLB-subject argument may not be as strong in South Carolina as in other states.

Another reason for selecting the subject area of science for initial implementation is the fact that South Carolina has eliminated constructed-response items from its current science assessments at elementary and middle school/junior high school levels. Thus, challenges related to capturing student responses to open-ended items and decisions regarding whether these responses would be scored by trained raters or through artificial intelligence, as well as issues related to turn-around time for reporting results, could be avoided in the initial implementation.

Additionally, the advice received from the Expert Panel supported starting with a lower-stakes subject area. As innovative item types, such as simulations and virtual experiments, are deemed appropriate for inclusion in South Carolina's science assessments, they could be implemented in a prudent manner.

WHAT GRADE?

Grade 7

Rationale: As a result of the implementation of South Carolina's Act 254, pre-high school science assessments are now census tests at grades 4 and 7. Thus, initial implementation of computerized testing could most easily take place at one of these census grades. Grade 7 is recommended, since students at that level will likely have had more experience using computers in school than students in grade 4 and would likely be less troubled if malfunctions occurred during testing.

WHAT TEST DESIGN?

Summative CBT and Formative Assessments

Rationale: It is recommended that the initial test should be a summative CBT for Grade 7 science and should be accompanied by a computerized formative test for Grade 7 science. The Grade 7 formative science assessment could be a single formative system provided by the state or one of the Grade 7 formative science assessments from the state-approved list, with the stipulation that districts must administer one of the assessments from that list. If this approach is taken, students will receive feedback about their academic strengths and weaknesses well in advance of the spring summative test so that they can receive targeted instruction on South Carolina's content standards, portions of which will be measured on the spring summative test. Thus, the link between assessment and instruction will be forged.

Even though CAT is in the forefront of research at this time and could target a student's actual instructional level to a greater extent than could many summative tests, NCLB requirements, plus the potential need for an extensive increase in the state's item bank for a summative assessment, suggests that a state-developed summative CAT system is not the best way to begin computerized testing at the state level. Specifically, it would be more cost effective to import viable items from the state's existing summative assessment item bank into the Grade 7 science CBT than to develop the additional items that would be needed to support a CAT. The preferred number of items to assure a well-designed CAT is approximately 450 per grade and subject; thus, initial implementation of a CAT would very likely necessitate additional item development from the start. Therefore, CAT should not be considered a viable choice for summative assessment in the first phase of the program in South Carolina. A Grade 7 summative science CBT accompanied by a formative companion testing system will strengthen the link between assessment and instruction. However, a CAT for Grade 7 science that has been approved and is on the state list of formative assessments could be used as the companion to the summative CBT.

WHOM TO TEST?

State representative sample of 1,500 students and additional volunteer districts/schools

Rationale: As described in Components 2 and 3, for purposes of state and, eventually, federal reporting, it is imperative that student test scores from summative tests be comparable within a test administration irrespective of the mode of assessment and across years, especially when the mode of administration changes. Thus, comparability studies are needed, and students selected for the study must be representative of the state in terms of multiple demographic characteristics. To ensure reliable and valid results from the comparability study, however, some design controls should be put in place. For example, the mandatory sample of students should be tested either in both modes or as part of a matched sample. Selecting a sample of 1,500 students will likely provide a sufficient number of students who actually test in order to account for attrition. Additionally, it is recommended that volunteer districts/schools be encouraged to participate in the computerized assessment, so that they may gain familiarity with the system and, perhaps, receive expedited test results as an incentive to participate.

Throughout each phase of the computerized testing rollout plan, a comparability study should be conducted as part of the initial administration of each computerized test.

WHICH STUDENT POPULATIONS?

General education students only

Rationale: Some states have begun computerized testing initiatives by working with special populations; others have chosen to implement programs for the general population. Since special population students often have unique needs with regard to assessment, it is likely that unique test forms, which incorporate accommodations needed for these students, will need to be developed, as is the current practice in South Carolina. In contrast, constructing a computerized test for the general population of students at a particular grade calls for developing one test form. Again, with so much to be accomplished in a limited period of time, this study supports simplifying the work to be accomplished in the first phase of implementation.

Alternate Proposal for Phase 1

As a result of discussions that took place at the June 13, 2007, Advisory Committee meeting, an alternative proposal for the initial implementation phase has been included: begin with Grade 6 mathematics, rather than Grade 7 science, and then follow the same roll-out plan as described for science in Phase 2 and beyond.

Grade 6 mathematics was recommended by several committee members for the following reasons:

- Since mathematics content standards progress in complexity from year to year but address common strands of learning across years, the immediate return of mathematics test results would provide timely information for educators to use in planning their instruction for the current and the following school year. On the other hand, South Carolina's science standards for scientific inquiry address learning strands that are common across years, but the other science standards, such as those for life science, are topic-focused within each grade. Thus, specific areas of need for students identified within science would most likely be grade-specific, rather than knowledge that builds across grades. This point was made by several members of the Advisory Committee who recommended beginning with mathematics rather than science.
- Tools typically available for use on mathematics tests, such as calculators, formula charts, rulers, and protractors, would be available as part of the electronic assessment itself, rather than as separate devices/tools that may not be standardized across the state. The inclusion of these tools as an integral part of the computerized mathematics assessment will ensure a high degree of standardization in test administration across the state.
- Since information from South Carolina's own accountability system includes test results from all four core content areas to report school and district performance, no one subject area is viewed as being more "high-stakes" than another, which is not the case with the current components of NCLB's AYP. Currently only mathematics and reading scores are used in AYP calculations.

Phase Two

WHAT SUBJECTS?

Science and Mathematics

Rationale: Based on the successful implementation of Grade 7 science in Phase 1, if that subject area and grade is chosen, it is recommended that all remaining grades of science from Grades 3-8 plus high school exit-level and end-of-course science tests that are not already being delivered by computer be administered electronically. Please note that currently the science tests at Grades 3, 5, and 7 are administered to a representative sample of students at each of those grades and not to the entire student population. Assigning specific subject-area tests to specific students is possible in some computerized testing systems through a computer algorithm. Such an automated system would greatly decrease the burden of manually assigning and distributing specific paper tests to specific students, as is the case now.

Additionally, it is recommended that one grade of mathematics should be added in Phase 2, if the plan to implement science first is followed. Even though mathematics is an NCLB-required subject area, as well as a component of South Carolina's accountability system, a phased approach for implementing this subject should be considered to ensure that sufficient resources are available in South Carolina schools.

A key component of the successful implementation of computerized mathematics assessment is student familiarity and comfort with online tools, such as calculators, rulers, and formula charts. Even though our assumptions listed later in this section assume tutorials and practice tests, these components of a successful computerized assessment are critical in mathematics and warrant special attention.

What Grades?

Science at all grades and Grade 6 Mathematics

Rationale: This aggressive plan reflects a model of introducing one grade of a subject in its initial phase and adding all grades of the particular subject in the subsequent phase.

What Test Design?

Summative CBT and formative assessments for all tests to be implemented in Phase 2

Rationale: This study recommends that each new assessment administered electronically should be accompanied with a corresponding formative assessment. The rationale for this recommendation is delineated in the Phase 1 test design rationale.

Whom to Test?

State representative sample of 1,500 students for all new grades/subjects to determine comparability and voluntary participation of other districts/schools

Rationale: The same rationale for comparability studies needed for all added grades/subjects described above and throughout this study pertains to Phase 2, as with Phase 1. Involvement by volunteer schools and districts should be encouraged, and efforts must be undertaken to share positive experiences with others.

Which Student Populations?

General education students and those special education students who can take a computerized assessment using appropriate tools that are offered through the chosen test engine

Rationale: During the second phase of implementation, discussions should take place with appropriate groups regarding the inclusion of special education students in the computerized testing delivery system. If the community of special educators rallies behind this initiative, Phase 2 could see some implementation for these students on a small scale, if the appropriateness of their participation is documented in their Individualized Education Plan. At a minimum, Phase 2 discussions should lay the groundwork for Phase 3 implementation.

Phase Three

WHAT SUBJECTS?

Science and Mathematics at all grades; and English language arts at grade 6

Rationale: This recommendation supports the pattern of starting with one grade of a new subject area and fully implementing and carrying forward those subject areas begun in previous phases.

An additional consideration must be addressed, however, with the implementation of English language arts (ELA) assessments via computer. If written composition remains as part of the ELA test, policymakers must address two issues:

- Will the compositions be scored exclusively by trained human raters or will artificial intelligence be used to derive at least one of the scores for the composition, with any discrepancies being resolved by a human rater?
- If the compositions will continue to be scored exclusively by human raters, will the ELA test as a whole be administered earlier in the school year so that scores can be reported at the same time as the science and mathematics assessments, or will the ELA tests be redesigned as separate reading and writing assessments so that just the writing test can be administered earlier in the school year?
- If the ELA tests are split into separate assessments, will the reading test be redesigned to include only multiple-choice items and/or short constructed-response items that can be scored using artificial intelligence or by human readers who receive the responses electronically?

These decisions must be made unless South Carolina students, parents, and educators are willing to continue to receive ELA test results following the end of the school year. In any of the scenarios for change that are bulleted above, it can be expected that the assessments will need to be redesigned, and proficiency levels will need to be reset if the test is either reconfigured or administered earlier in the school year. In addition, the redesigned tests will likely need to be submitted to the U.S. Department of Education for approval.

Paths for Implementation

Based on the scenarios for implementation of computerized testing described above, many different paths may be chosen on the way to fully implementing online testing in terms of grade and subject combinations and the order of implementation. Rather than assert that there is one single, clear direction, consider the following options:

- no change – continue on current path, or
- choose computerized testing, with either science or mathematics as the first subject area to be converted to this format

NO CHANGE – CONTINUE ON CURRENT PATH

Maintain current paper-based assessment delivery system.

Choose Computerized Testing

Suggested Implementation Plan

Phase 1	• Science – one grade (grade 7)
Phase 2	• Science – all grades • Mathematics – one grade (grade 6)
Phase 3	• Science – all grades • Mathematics – all grades • English language arts – one grade (grade 6)
Phase 4	• New subject – one grade • Science – all grades • Mathematics – all grades • English language arts – all grades

Rationale: This aggressive plan will dramatically accelerate innovation in the schools; utilize the most appropriate, useful, diagnostic testing in the major subject areas in order to provide results more quickly; and allow teachers to make instructional decisions based on student test results. ***This study recommends that the implementation plan outlined above should be evaluated carefully and that high degrees of success in terms of state capacity, sufficient infrastructure, and adequate staffing should be evidenced before South Carolina moves to the next phase.***

Assumptions

The following are the assumptions underlying the above implementation plan.

- A phase does not necessarily correlate to one year and the time period for each phase does not need to be the same.
- Similar numbers of multiple-choice and constructed-response items (constructed-response and extended-response items scored by human readers) that are found in the current state tests.

- A robust bank of test items exist and these items can be converted from paper/pencil format to computerized display.
- Tutorial(s) will be developed for each grade and subject
- Electronic surveys of examinees' experiences with computers will be administered
- Practice test(s) for each grade and subject will be developed and made available in the fall prior to each new computerized test.
- Accommodations within vendor package (e.g., highlighting, strike-through, text-to-speech, calculator, ruler, protractor) will be provided
- A minimum testing window of three weeks, which may increase with volume of tests/students, will be needed
- While South Carolina may choose to build its own formative assessment item bank for schools to use, *no costs for such development have been included in this plan.*
- Participation rate:
  - Per subject introduced (science, mathematics, and then English language arts):
    Phase 1 of subject = comparability sample plus 10% voluntary,
    Phase 2 of subject = comparability sample plus 50% voluntary,
    Phase 3 of subject = 90% mandatory. This formula is repeated for each subject introduced
  - Phase 1 10% participation: assumes computers at schools meet system requirements
  - Schools volunteering would test entire grade (not just selected students)

- Comparability studies would be conducted for each new subject and each newly added grade
  - Voluntary participation (Phase 1 10%) plus representative 1,500 sample for comparability study
- Formula can become more or less aggressive by pulling forward or pushing out integration of additional subjects—high school graduation retesters; end-of-course non-NCLB, end-of-course NCLB, writing)

IMPLEMENTATION PLAN COSTS FOR BUDGETARY PURPOSES ONLY

Reasonable estimates are provided below for the implementation plan shown above. Due to the dynamic nature of the testing industry, the speed of implementation that South Carolina chooses, the possible volume of students participating, and the changing costs of technology, the below numbers are estimates only. *Computerized testing will not result in significant cost savings.*

NOTE: Costs for bringing the state, districts, and schools to needed capacity for administering computerized tests are included in Components 7 & 9 and are NOT reflected in the implementation costs delineated below.

1. Phase 1 – $1 million
   - See Assumptions
2. Phase 2 – $4 million
   - See Assumptions
3. Phase 3 – $12/per test/per student
   - Must assume significant volume of students participating (90% at year 3 for each subject tested)

4. Phase 4 – $11/per test/per student

5. Phase 5 – $ 11/per test/per student

Other Recommendations

General

- Assessment, technology, and policy decision makers must present a coordinated effort with commitment and support from their staffs in all areas and at all levels. Successful implementation will require the best efforts of all.

- The South Carolina Technical Advisory Committee should be actively involved in the planning and implementation of computerized testing.

- The reauthorization of No Child Left Behind could have implications for large-scale assessment programs, including computerized testing, and should be tracked carefully as rollout plans for computerized testing are developed for South Carolina.

- Consider utilizing the Council of Chief State School Officers (CCSSO) Online Computer-Based Decision Making Tool. The tool is intended to assist state department personnel in making decisions and in implementing plans to move from a paper-and-pencil testing environment to a computer-based testing environment.

- Many states are currently using electronic tests as major components of their assessment programs. These early implementers can provide important lessons learned that should be used to evaluate what was effective, what was not, and why.

- Effective back-up plans should be put in place in case of catastrophe.

Assessment-related

- Test results are of greatest value if returned to teachers promptly after testing. The needs of the teachers could be accommodated without compromising the accountability function with *two-stage reporting,* since preliminary data that are useful to teachers do not require final scaling or complete normative data. This first wave of reporting could be based on multiple-choice items only or could include constructed response data if some form of automated scoring were used for the preliminary results.

- Formative assessments for classroom use could include constructed-response items that are locally scored by the teacher. Constructed-response data are of limited diagnostic value unless the teacher can see the actual student response, which is permitted in some state assessment programs. In these programs constructed-response items and scoring rubrics are released publicly following test administration, are not re-used, and students' responses to constructed responses are either returned to the school or schools are permitted to copy the response prior to its being sent for scoring.

- Comparability studies should be conducted during the initial year of administering each new test via computer. While running parallel electronic and paper systems, a matched-sample comparability study should be a routine part of the analysis.

- Electronic delivery has the potential to provide a variety of accommodations for special needs students, and students who can benefit from the use of these accommodations should be assessed via computer as soon as feasible.

Technology

- A realistic try-out of the infrastructure should be conducted at each school to ensure that the infrastructure is adequate to handle the test itself, as well as the numbers of students who will be testing. This try-out should include consideration of time of day, day of week, and competing uses, and the number and capacity of the computer stations.
- To ensure equity of access for all students in the state, technology content standards for students should be defined and included in the curriculum.
- Specific technology content standards for both students and educators should be developed. Mastery of those standards for educators should be measured in initial licensure examinations or as a required component of educators' continuing education.
- Ongoing professional development in technology should be provided to teachers to ensure that they can lead their students in these emerging areas.

Investment

- Depending on the implementation schedule chosen, a significant investment in both infrastructure and staff will be required to ensure all schools and districts have equitable access. This investment should be sufficiently funded to lead to success.

COMPUTERIZED ADAPTIVE TESTING

One important variation of electronic testing is *computerized adaptive testing*[1] (CAT). The primary advantage of CAT is that it can be used to efficiently produce a precise measurement of a person's location along a construct continuum. By targeting the person's location at each step, the number of items required to obtain a precise estimate of status can be dramatically lower than with a fixed form. Originally, this efficiency was intended to minimize testing time, but it can be used to collect more diagnostic information as well.

The CAT solution generally presumes the underlying scale or construct to be measured is continuous, unidimensional, and unbounded. This conceptualization is somewhat at odds with the reliance of large-scale assessment on grade-level, content-standards-linked items. CAT and standards-based assessments are not irreconcilable. A CAT algorithm could administer items in large-scale assessment in an efficient manner, with the restriction that only on-grade material would be presented.

In contrast to a fixed form where each student takes the same items in the same order, a pure CAT algorithm selects the *optimal* item to present to each student based on the responses to the preceding items. There are variations on both formats. For security reasons, *fixed form CBT* may use several scrambled or parallel versions. For logistical and administrative reasons, some *CAT* algorithms present small, fixed blocks of items, or testlets, between decision points rather than single items with decision points after each item.

Psychometrically for its original purpose, CAT will always be more *efficient*, i.e., require fewer items, than a fixed form to achieve a desired level of precision. To take an extreme

[1] To date, NCLB has been reluctant to accept CAT assessments for its purposes. Specifically, it requires all NCLB testing be done with on-grade, standards-based items. It also has tended to favor fixed form assessments over individually customized tests. These issues will be discussed in more detail in Component 3.

example, if the items from a single fixed form comprise the entire item pool available to a CAT assessment, there are few cases for which the CAT algorithm will find it productive to administer all the items to any individual. So, for most cases, the CAT will be shorter for the same level of precision even with a very small item pool. For very high or very low achieving students, the fixed form may, in fact, never reach a desirable level of precision. It depends entirely on whether the item pool consists of items that match the achievement level of the particular student.

In spite of the generally accepted psychometric advantages, CAT is often viewed with skepticism by the public and policy makers. The concerns focus on a few issues:

1. Fairness of comparing student scores derived from different selections of items.
2. Selection of items to match the blueprint of a standards-based assessment.
3. Differences in effective testing-taking strategies between adaptive and fixed forms.

The question of the *fairness* of an assessment that does not administer exactly the same items to every student is not an issue for psychometricians. On the contrary, psychometricians might question if it is *fair* to administer items for which the student has almost no chance of succeeding.

The basic premise underlying all educational assessments is that all items in the pool, which are available for use on a test, pertain to the same unidimensional construct. With CAT, the item difficulties have been established before the items are placed into the pool and the item difficulties are taken into consideration when scoring a test; hence any form drawn from the pool will be automatically equated to every other possible form. The measurement model adjusts for both differences in test length and in the distribution of item difficulties.

How efficiently and precisely the estimation is accomplished depends almost exclusively on how well the items are matched to the individual. This matching of the test to the individual is the essence of CAT.

Although differences in item difficulty and test length can be easily dealt with by the CAT models, one might still argue that the content sampling in CAT could advantage (or disadvantage) a student who, for whatever reason, happens to be more (or less) familiar with the content than is the typical student. While this is not an issue when considering the total population, a student with unusual experience or a class with unusual instruction could be affected positively or negatively by the specific choice of items.

However, because any form is just a sample or subset of the content domain, students can always argue they would have done better with a different selection of items. This argument is no more valid as it pertains to a customized CAT (*I might have done better if I had gotten the items that my friend got)* or a fixed-form paper & pencil test (*I might have done better if I had gotten last year's form.)* Properly implemented, a CAT is capable of supporting a fair decision about each student's correct performance level assignment.

Possibly the best defense against the unfairness argument is to ensure that the CAT form matches the *blueprint* of test specifications. Traditionally, the blueprint has included all the details needed to construct the test forms. Total number of items, distribution of items across content standards, appropriate difficulty levels, acceptable item formats and the like would all be described. For CAT, these constraints are implemented through the available item pool, the item selection algorithm, and the termination rules. Historically, CAT item selection algorithms have focused primarily on the difficulty level of the item[2] without concern for any other characteristics

[2] For Rasch models, the item is optimal if the difficulty level matches the current ability estimate. For more complex IRT models, the optimization is more complex but has the same intent of maximizing the amount of information expected from the next response.

like content, standard, grade-level, format, exposure, security, or taxonomic level[3]. Much of the recent CAT literature considers strategies to deal with the management of issues related to the balancing all constraints (Luecht, 2005a; Drasgow, Luecht, and Bennett, 2006; Glas & van der Linden, 2000; van der Linden, 1998, 2000).

Considerable research has been done on alternative test delivery models that might be considered by South Carolina. A number of certification and licensure organizations, as well as at least one state (Oregon), that originally considered CAT instead decided to adopt the alternative test delivery model *ca-MST*, computer-adaptive multistage testing (Luecht & Nungester, 1998; Luecht & Nungester, 2000; Luecht, 2000; Luecht, 2004).

A ca-MST is adaptive, but employs pre-constructed modules arranged in self-adapting packages called "panels." The ca-MST is similar to but more carefully controlled than adaptive testlet models. The ca-MST approach is almost as efficient as an item-level CAT, but offers greater quality control over content and other test quality features than CAT, simplifies the real-time scoring and routing mechanisms needed, and provides powerful item-exposure controls.

Van der Linden (2000) has proposed a method he refers to as shadow tests using integer programming with all the constraints of the blueprint and the situation. After each item is administered and score, the item selection routine selects an optimal test based on the test blueprint rather than a single item. The single item that is administered is the optimal item from optimal test. As a result, when testing terminates, the student will have taken the optimal test given the responses and that test must necessarily be in compliance with the blueprint.

The requirement that no out-of-level or off-grade item can be administered is one way of constraining the item selection. When the intent of CAT is to determine the person's location,

[3] Because a unidimensional construct is being measured, these additional characteristics are extraneous.

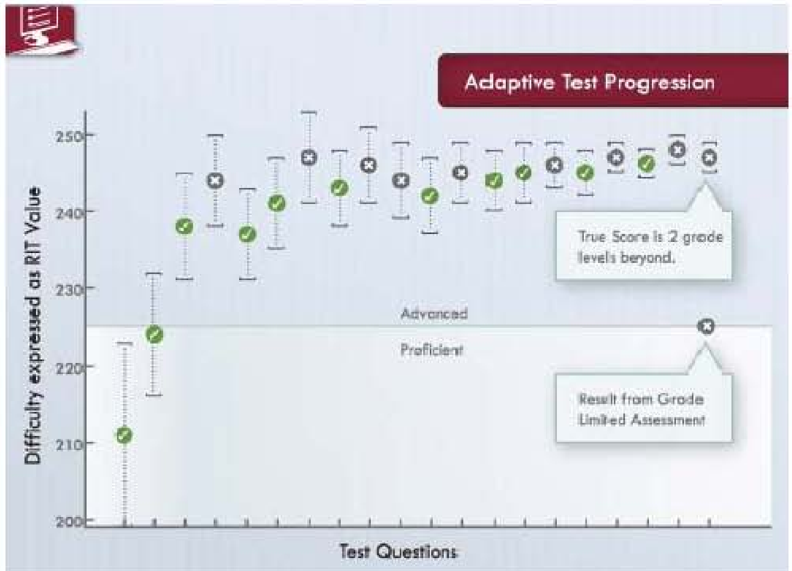
wherever that might be, the distinction as on- or off-grade is not germane[4]. The student's location was to be estimated as accurately as possible, even if that meant drawing items officially intended for higher or lower grades. There is little to be gained psychometrically from administering an on-grade item if the student will almost certainly get it right or almost certainly get it wrong.

Allan Olson, co-founder and former Chief Academic Officer of Northwest Evaluation Association, testified before Congress, "To be able to measure achievement for these students [well beyond or behind grade level] requires a measurement scale that goes beyond grade-level testing and identifies what students know across the many strands of knowledge that a student needs to know to be identified as a proficient." (Olson, 2007). In his testimony, Olson was not arguing that CAT be admitted under the current legislation, but that new legislation include true growth models supported by CAT. It is in the context of measuring growth that beyond-grade-level testing is required.

When the objective is to determine proficiency at grade-level, as with NCLB, rather than location on a broad developmental continuum, on-grade testing is appropriate and easily ensured: *if there are no off-grade items in the pool, the CAT algorithm cannot go off-grade*. From a CAT perspective, high and low ability students will have biased estimates of ability (overestimates for low students; underestimates for high students) but from a standards-based assessment perspective where the intent is to classify the students into the appropriate performance level, CAT should still be more efficient than a fixed form.

[4] What is germane to any assessment is that the person not be tested unfairly. No one should be tested with items on content that has not been presented or on skills that have not been reinforced. This is fundamental to any assessment and is an especially nagging issue for any assessment that relies on longitudinal scales.

Figure CAT.1



The graphic above, taken from Olson's testimony (2007), illustrates the potential bias for a high-achieving student. In this case, the CAT resulted in a measure for the person about 20 scale score points higher (which Olson suggests is two grade levels) than if the assessment were restricted to grade-level[5].

In general, the item selection will be less than optimal, in the traditional sense of maximizing information, and will require more items when additional constraints and requirements are added. While this situation may not represent the most efficient CAT implementation, a CAT can still be used effectively to estimate the student's location relative to the performance standards levels and to provide relevant grade-level content standard results in the context of standards-based assessment.

[5] This example makes its point about potential bias but somewhat overstates the problem. The grade-level assessment is not necessarily bounded by the Proficient/Advanced border.

Item Pool Size

It is often stated that CAT requires a very large item bank. It is undoubtedly true that to realize all the potential power and elegance of CAT, a large bank is useful. To use CAT efficiently, there needs to be sufficient items for each standard being assessed and enough items at the extremes of the difficulty continuum to permit the selection of items with appropriate attributes. While these limitations are present with fixed form development as well, they are more evident in the CAT environment.

It is also true that ongoing item development and tryout are necessary with any assessment to replenish the item supply when items are released, overexposed, compromised, or outdated. CAT may interact with the exposure issue because, on the one hand, fewer students see any one item while, on the other hand, longer testing windows imply more items are needed to mitigate the effect of possible student conversations about the test contents.

Halkitix (1998) investigated the precision of a computerized adaptive test (CAT) with a limited item pool using test results from 4,494 nursing students. Regardless of the item pool size, CAT provided greater precision in measurement with a smaller number of items administered even when the choice of items was limited. However, with a limited pool, CAT failed to achieve one of its important goals, which is equal precision along the entire ability continuum.

With a limited bank, the measurement at the extremes of ability will typically be biased because the item selection algorithm would not be permitted to move as far as it would like. The same effect would occur with a fixed form operating from the same item pool as well. The fixed form will generally use more items and result in larger standard errors and biases at the extremes.

Using an item pool of limited size is similar to using a pool that is restricted to a specific grade level. The CAT will not perform as efficiently as it might in the optimal situation but it will use fewer items to achieve an acceptable standard error than will the fixed form version in a similar environment.

The strongest force driving a large bank for CAT is test security. With fewer students being testing at the same moment, students will have more opportunity to share information than with the conventional, group-administered fixed form. This forces a large pool to minimize overlap.

Item Review

When questioned about their experience taking an electronic test, examinees at all levels generally feel strongly that they be allowed to review and alter their responses to earlier items. While item review and change are possible with electronic testing in general and CAT in particular, they typically are not permitted with a pure CAT.

The initial argument against review was that the items selected were *optimal* for the responses given. Allowing students to alter their responses would ensure the path taken was not optimal and thus led to a loss of information and biased estimates (Lord, 1983).

The argument that examinees should not be permitted to review and change responses on a CAT because it leads to biased estimates[6] appears specious, and perhaps pernicious, in the case of legitimate errors. If the examinee inadvertently selected an incorrect response through misreading, inattentiveness, or distraction, the corrected response will always better represent the true ability than would the careless error. An ability estimated from the corrected data will be more accurate, valid, and optimal than one based on bad data, regardless of what the information function says. If the examinee makes and corrects careless errors, the path through the items will

[6] Because bias is inversely related to information, a loss of information caused by allowing the CAT to drift off target will lead to an increase in bias (Lord, 1983).

be less efficient than it might be but that does not seem sufficient reason not to obtain the best data possible.

A common argument is that a clever examinee might exploit some aspects of the CAT algorithm to artificially inflate the estimate of ability beyond what is deserved. Several clever researchers have devised schemes for how one might subvert the process in this way.

Wainer Strategy

Using the Wainer Strategy (Wainer, 1993; Wise, 1996), the examinee attempts to obtain a higher ability estimate by forcing an easier test for which all the initially wrong answers can be changed to right at the finish. This is accomplished by deliberately missing items for which the examinee is certain of the correct answer and correcting the responses after the CAT algorithm satisfies its termination rule.

Simulation studies have demonstrated that the strategy, perfectly executed, can result in an overestimate but that the strategy is high risk (Gershon & Bergstrom, 1995). An overestimate happened only when every revised item was answered correctly; even one mistake resulted in an estimate below the true ability. This review could also result in very large standard errors of measurement. The adverse effects can be mitigated by the methods used to select items, to estimate ability, and to terminate testing (Bowles & Pommerich, 2001).

Kingsbury Strategy

The Kingsbury Strategy depends on the examinee recognizing when the item presented is easier than the preceding one (Green, Bock, Humphreys, Linn, & Reckase, 1984; Kingsbury, 1996.) In this situation, the astute examinee will know the previous response was incorrect and will change it if review is allowed. Successfully implementing the strategy depends on the examinee being aware of how the CAT selects items and, more problematically, being able to determine accurately the direction of the adjustment in item difficulty.

Wise, Finney, Enders, Freeman, and Severance (1999) found that when items differ by a half logit or less, examinees were right one-half of the time (essentially, a coin toss) in judging which item was more difficult. In contrast, with more than one-half logit difference, they were right about three times in four. These errors rates suggest the strategy is unlikely to have a significant benefit.

Simulation studies have supported this impression. Depending on exactly how the strategy was implemented and how the examinee's ability to discriminate differences in item difficulty was parameterized, simulations have shown biases of 0.11 logits (Kingsbury, 1996) and 0.01 and -0.03 logits (Wise et al, 1999).

Again, the strategy for changing previous answers can be effectively defeated by minor modifications to the item selection algorithm with minimal harm to efficiency (Bowles & Pommerich, 2001; Papanastasiou, 2002).

There is a real concern that examinees might try to subvert the process by adopting the Wainer, Kingsbury, or other strategies. These often place the CAT assessment in the unfamiliar position of having administered a test that was poorly matched to the examinee. While the effects can be mitigated by the item selection and ability estimation algorithms, the most effective defense would involve the rule for terminating testing.

If the effect of the item review[7] is a substantial change in the ability estimate or increase in the standard error, testing should continue until the estimate is stable and the precision is satisfactory. This can be inefficient in the sense of requiring a longer test, but that should be of

[7] With electronic testing, it is relatively straightforward to detect the behaviors associated with either of the strategies described.

less concern than obtaining good data. Subversive tactics may become less appealing to examinees when they discover the tactics lead to longer test sessions with little or no benefit.

From the opposite perspective, there are important reasons to allow changes. First, examinees have a strong preference for being allowed to review previous items at any time during the testing session, as with paper and pencil. Second, eliminating careless errors will more accurately reflect the student's true level of proficiency. Either reason alone is a persuasive argument for allowing item review.

Summary

There is little reason to consider CAT if NCLB excludes growth models and if it requires proficiency classifications be based on a common core of on-grade items.

- CAT is more efficient than a fixed form assessment to achieve the same ends.
- CAT can match a designated test blueprint if properly managed. This includes using only grade-level items and adequate sampling within content standards to provide effective diagnostic feedback.
- CAT can detect and mitigate strategies for over-exposure cheating, guessing, and manipulating the system with appropriate item selection and termination rules.
- Appropriate item selection and termination rules may imply alternative test delivery models, such as computer-adaptive, multi-stage testing (*ca-MST*) or *shadow tests*.
- CAT may require significant start-up efforts for item development to provide adequate numbers of items across the continuum and across the content standards.
- CAT will require effective communication for educators and the public to accept test results based on individually customized tests.

COMPUTERIZED ADAPTIVE TESTING – LINKS AND REFERENCES

Bowles, R. & Pommerich, M. (2001). *An examination of item review on a CAT using the specific information item section algorithm*. Paper presented at the Annual meeting of the National Council on Measurement in Education, Seattle, WA.

Drasgow, F.; Luecht, R. M.; & Bennett, R. E., (2006). Technology and testing. In R. L. Brennan, *Educational Measurement, Fourth Edition* (pp. 471–515). Westport, CT: American Council on Education and Praeger Publishers.

Gershon, R. & Bergstrom, B. (1995). *Does cheating on CAT pay: NOT!* (ERIC Document Reproduction Number TM024692).

Glas, C. A. W., & van der Linden, W. J. (2003). Computerized adaptive testing with item clones. *Applied Psychological Measurement, 27*, 247-261.

Green, Bock, Humphreys, Linn, & Reckase, 1984. Technical Guidelines For Assessing Computerized Adaptive Tests. *Journal of Educational Measurement 21 (4), 347–360.*

Halkitix, P., (1998). The Effect of Item Pool Restriction on the Precision of Ability Measurement for a Rasch-Based CAT: Comparisons to Traditional Fixed Length Examinations. *Journal of Outcome Measurement*, *v2 n2 p97–122*.

Kingsbury, G. (1996) *Item review and adaptive testing*. Paper presented at National Council on Measurement in Education. New York, NY.

Lord, F. (1983) Statistical bias in maximum likelihood estimators of item parameters. *Psychometrika*, 48, No. 3.

Luecht, R. M. (2005a). Operational Issues in Computer-Based Testing. In D. Bartrum and R. Hambleton (Eds). *Computer-Based Testing and the Internet*. New York, Wiley & Sons Publishing.

Olson, A. (2007) *Testimony concerning the measurement of student achievement growth*. Retrieved May 8, 2007 http://edlabor.house.gov/testimony/032107AllanOlsontestimony.pdf

Papanastasiou, E. C. (2002). *A "Rearrangement Procedure" for scoring adaptive tests with review options*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

van der Linden, W. (1998). Optimal assembly of psychological and educational tests. *Applied Psychological Measurement*, 22, 195–211.

van der Linden, W. (2000). Constrained adaptive testing with shadow tests. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: theory and practice*. (pp. 27–52). Boston, MA: Kluwer Academic Publishers.

Wainer, H. 1993. Some practical considerations when converting a linearly administered test to an adaptive format. *Educational measurement, issues and practice*, 12, 15–20.

Wise, S. L. (1996). *A critical analysis of the arguments for and against item review in computerized adaptive testing*. Paper presented at the annual meeting of the National Conference on Measurement in Education, New York, NY.

Wise, S. L., Finney, S. J., Enders, C. K., Freeman, S. A. & Severance, D. D. (1999). Examinee judgments of changes in item difficulty: Implications for item review in computerized adaptive testing. *Applied Measurement in Education*, 12, 185–198.

SURVEY METHODOLOGY

One of the sources of information for the South Carolina Feasibility Study was a Web survey of the districts and schools within the state. The Web survey was used to collect current information about a variety of technology issues within districts and schools in South Carolina that are critical to assessing the feasibility of, and planning for, conversion to computerized testing. The survey provided information that was incorporated throughout the 14 major components of the Feasibility Study Report. This section of the final report documents the actions that were taken to create the Web survey, to administer the survey to South Carolina districts and schools, to analyze the survey responses, and to generate the survey results.

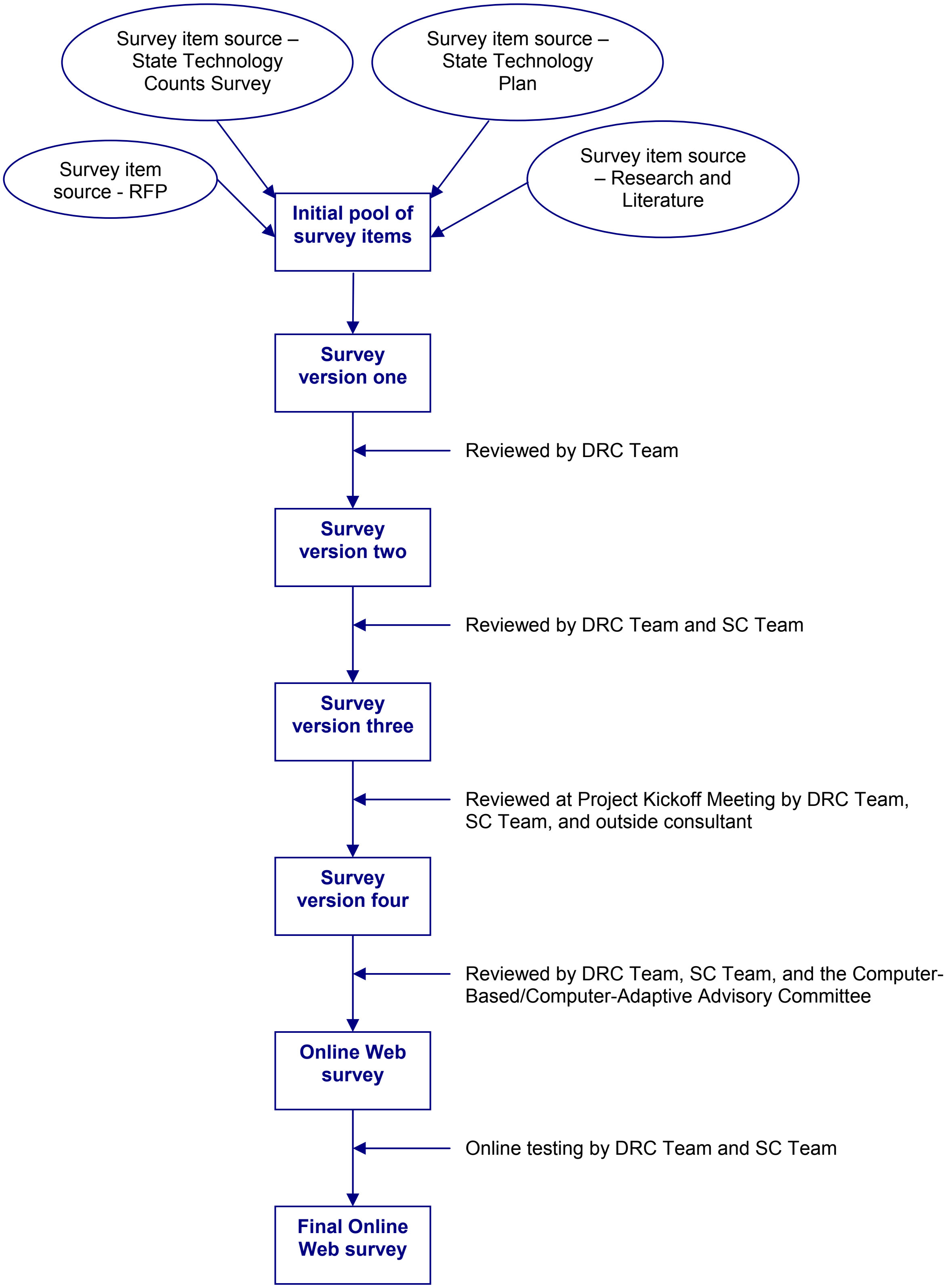
Content Creation Process

The request for proposal (RFP) for the Feasibility Study included a list of the initial topic areas for the Web survey content. These topic areas were researched and expanded upon to create a list of potential survey items. Additionally, survey topic areas were gleaned from the fourteen report components. Other resources included the State Technology Counts Survey and the State Technology Plan. The survey topics and items for survey version one were compiled from these resource areas. Survey version one was reviewed by the DRC Feasibility Study team members (DRC Team) who are content experts and who were tasked with writing the feasibility study. Survey items were refined, added, and reorganized. In addition to the survey draft, a content guide/map was created as an internal working document to ensure that all the survey topics listed in the RFP were sufficiently covered by one or more survey items. After this review process, survey version two was created.

Survey version two was sent to the South Carolina Feasibility Study team (SC Team) for review and was reviewed once again by members of the DRC Team. The SC Team discussed their revisions to survey version two via a conference call with DRC. Comments from the phone call were used to create survey version three. Survey version three was discussed with the attendees of the Project Kick-off Meeting, January 29 and 30, 2007 in Columbia, SC. Contributing to the review were SC Team members, DRC Team members, and an external project consultant. Feedback from the meeting was used to create survey version four. Survey version four was made available to the SC Team and DRC Team for a final review before an online version was developed. The Computer-Based/Computer-Adaptive Advisory Committee also reviewed and provided feedback about survey version four at their February 8, 2007 meeting.

The survey items from survey version four, along with the feedback from SC Team members, Advisory Committee members, and DRC Team members were then used to create the SC Feasibility Study Web Survey. A "Frequently Asked Questions" (FAQ) page was created, reviewed, and added to the Web survey to provide useful information (e.g., definitions, submission date, etc.) to the participants. The DRC and SC Teams tested the online version. Minor changes were made after testing was complete, the final survey content was agreed upon, and the content was locked into the Web survey system. The survey content process flowchart is displayed in Figure S-1 and the final Web survey content can be found in Appendix A.

Figure S.1: Survey Content Creation Process Flowchart



Communications and Administration Chronology

Multiple communication modes and channels were used to notify districts and schools about the survey for the Feasibility Study. Both the SC and DRC Teams were involved in crafting communications and publicizing the importance of the survey leading up to the administration window of February 21 through March 7, 2007.

1. On February 15, 2007, Dr. Teri Siskind and Elizabeth Jones notified attendees at the Instructional Leaders Roundtable meeting about the survey. Attendees received advanced copies of the survey content and a questions-and-answers document about the survey process. Dr. Siskind and Ms. Jones responded to questions about the survey from attendees during the meeting.
2. On February 16, 2007, the SC Team publicized the upcoming survey via an e-mail from Dr. JoAnne Anderson and Dr. Jim Rex. DRC provided suggested content and phrasing for this e-mail to the department. The SC Team modified it to reflect the message that Dr. Anderson and Dr. Rex wanted emphasized.
3. On February 16, 2007, DRC sent a pre-notification e-mail to all District Test Coordinators. All District Superintendents were sent a copy of the e-mail. The e-mail provided information about the purpose of the study, details about the survey process, the timeline for survey completion, and their role in the survey process. Additionally, a survey worksheet was attached that districts could print and use to begin gathering the information needed to respond to the survey questions.
4. On February 20 and 21, 2007, Elizabeth Jones attended the District Test Coordinator (DTC) workshop. Ms. Jones communicated with District Test Coordinators, emphasizing the importance for all districts and their schools to complete the survey. She also addressed questions related to the survey.

5. On February 21, 2007, DRC sent the invitation e-mail to each District Test Coordinator. The e-mail included Web survey access instructions for the district portion of the survey, as well as instructions for completion of the school portion of the survey. The survey worksheet was included as an attachment to the e-mail. The District Test Coordinators were instructed to pass along the survey instructions for schools to the School Test Coordinators and/or other appropriate school staff that may be best suited to gather the answers.
6. On February 28, 2007, at the request of some school districts, DRC notified District Test Coordinators that alternative schools had been added to the survey process. Several school districts provided input that their alternative schools administer state testing programs, so the SC and DRC Teams concluded it was important to add them to the process. If a district did not have an alternative school, the District Test Coordinator did not receive an e-mail.
7. On February 28, 2007, DRC sent a reminder e-mail to all District Test Coordinators. The message reiterated the importance of completing the survey by the March 7, 2007 due date and included instructions for accessing the Web survey. The District Test Coordinators were instructed to pass along the reminder to the School Test Coordinators and/or other appropriate school staff.
8. On March 2, 2007, in response to requests from districts, DRC notified District Test Coordinators that adult education centers had been added to the survey process. As was the case with alternative schools, several school districts provided input that their adult education centers administer state testing programs. If a district did not have an adult education center, the District Test Coordinator was not notified of the addition.

9. On March 6, 2007, the SC Team sent a reminder about the survey to members of the Instructional Leaders Roundtable. The reminder emphasized the importance of providing information regarding computerized testing via the survey.

10. After reviewing the districts and schools that responded to the survey by March 7, 2007, the SC and DRC Teams determined that the survey administration window would remain open until 5:00 p.m. Eastern time on March 12, 2007. On March 9, 2007, the Superintendents of seven districts that submitted no or few surveys were notified of the new deadline for completion of the survey. They were encouraged to contact the District Test Coordinator in order to expedite completion of the survey for their district and schools.

Web Survey Participant Demographics

Surveys were collected for 906 schools, 20 adult education centers, 22 alternative schools, and 86 districts. In one situation, data were submitted at the school level but not at the overall district level (i.e., data are available at the *school level* for 87 districts and at the *district level* for 86 districts). Table S-1 provides a breakdown, by district, of the number of districts, schools, adult education centers, and alternative schools that responded to the Web survey.

Table S.1: District Responses
(listed by number of responses/total number possible)

District	District	School	Adult Education Center	Alternative School
Abbeville	1/1	9/10	0/0	0/1
Aiken	1/1	37/41	0/6	0/1
Allendale	1/1	4/4	1/1	1/1
Anderson 1	1/1	13/14	0/0	1/1
Anderson 2	1/1	7/7	0/1	0/0
Anderson 3	1/1	4/4	0/0	0/0
Anderson 4	1/1	5/6	0/0	0/0
Anderson 5	1/1	15/16	1/1	1/1
Bamberg 1	1/1	4/4	0/1	0/0
Bamberg 2	1/1	2/3	0/1	1/1
Barnwell 19	1/1	3/3	0/1	0/1
Barnwell 29	1/1	3/3	0/1	0/1
Barnwell 45	1/1	4/4	0/1	0/1
Beaufort	1/1	16/28	0/2	0/1
Berkeley	1/1	36/36	0/2	1/1
Calhoun	1/1	4/4	0/1	0/1
Charleston	1/1	76/84	1/2	1/2
Cherokee	1/1	19/19	1/1	0/1
Chester	1/1	13/13	1/1	0/1
Chesterfield	1/1	15/16	1/1	0/1
Clarendon 1	1/1	3/3	0/0	0/1
Clarendon 2	1/1	6/6	0/1	0/1
Clarendon 3	1/1	1/3	0/0	0/1
Colleton	1/1	11/12	0/1	1/1
Darlington	1/1	21/23	0/1	0/1
Dillon 1	1/1	2/3	0/0	0/1
Dillon 2	1/1	6/6	1/1	0/0
Dillon 3	1/1	4/4	0/0	0/0
Dorchester 2	1/1	16/19	1/2	1/1

Table S.1 (continued): District Responses (listed by number of responses/total number possible)

District	District	School	Adult Education Center	Alternative School
Dorchester 4	1/1	5/5	0/0	0/1
Edgefield	1/1	3/9	0/1	0/1
Fairfield	1/1	5/9	0/1	0/1
Florence 1	1/1	16/21	1/1	0/1
Florence 2	1/1	1/2	0/1	0/1
Florence 3	1/1	8/8	1/1	1/1
Florence 4	1/1	3/3	0/1	0/1
Florence 5	1/1	3/3	0/1	0/1
Georgetown	1/1	15/17	0/1	1/1
Greenville	0/1	6/96	1/1	0/3
Greenwood 50	1/1	14/14	0/1	0/1
Greenwood 51	1/1	3/3	0/0	0/1
Greenwood 52	1/1	3/4	0/0	0/1
Hampton 1	1/1	0/7	0/1	0/1
Hampton 2	1/1	1/3	0/1	0/1
Horry	1/1	42/47	0/2	0/1
Jasper	1/1	1/4	0/1	0/1
Kershaw	1/1	18/18	0/1	1/1
Lancaster	1/1	18/19	0/1	0/1
Laurens 55	1/1	11/11	0/1	1/1
Laurens 56	1/1	7/7	0/1	0/1
Lee	1/1	7/9	0/1	1/1
Lexington 1	1/1	21/21	0/1	1/1
Lexington 2	1/1	16/16	0/1	0/1
Lexington 3	1/1	4/4	0/1	0/1
Lexington 4	1/1	4/6	0/1	0/0
Lexington 5	1/1	19/19	1/1	1/1
McCormick	1/1	3/3	1/1	0/1
Marion 1	1/1	4/4	0/1	0/1

Table S.1 (continued): District Responses
(listed by number of responses/total number possible)

District	District	School	Adult Education Center	Alternative School
Marion 2	1/1	4/5	1/1	0/0
Marion 7	1/1	3/3	0/1	0/0
Marlboro	1/1	5/9	0/1	0/1
Newberry	1/1	13/14	1/2	1/1
Oconee	1/1	19/21	1/1	1/1
Orangeburg 3	1/1	6/6	0/2	0/2
Orangeburg 4	1/1	8/8	0/2	0/1
Orangeburg 5	1/1	13/14	0/3	0/1
Pickens	1/1	23/25	1/2	1/1
Richland 1	1/1	10/51	0/1	0/1
Richland 2	1/1	16/25	0/1	0/1
Saluda	1/1	5/5	0/1	0/1
Spartanburg 1	1/1	6/9	0/1	1/1
Spartanburg 2	1/1	10/13	0/0	0/0
Spartanburg 3	1/1	7/7	0/0	0/1
Spartanburg 4	1/1	4/4	0/0	0/1
Spartanburg 5	1/1	9/10	0/0	0/0
Spartanburg 6	1/1	14/14	0/0	0/1
Spartanburg 7	1/1	13/14	1/1	0/2
Sumter 2	1/1	15/15	0/0	0/1
Sumter 17	1/1	9/11	1/1	0/1
Union	1/1	10/11	0/1	0/1
Williamsburg	1/1	11/16	0/1	0/1
York 1	1/1	6/8	0/1	0/1
York 2	1/1	9/9	0/1	0/1
York 3 (Rock Hill)	1/1	20/26	1/1	3/3
York 4	1/1	9/10	0/1	0/1

Table S.1 (continued): District Responses (listed by number of responses/total number possible)

District	District	School	Adult Education Center	Alternative School
Felton Lab	0/1*	0/1*	0/0	0/0
John De La Howe	0/1	0/2	0/0	0/0
Wil Lou Gray	0/1	0/1	0/0	0/0
School for Deaf	0/1	0/5	0/1	0/0
Department of Juvenile Justice	1/1	3/17	0/0	0/0
Palmetto Unified Schools	1/1	6/20	0/0	0/0
Thornwell	0/1	0/1	0/0	0/0
Total	**86/92**	**906/1200**	**20/85**	**22/81**

* This school printed and mailed a completed survey to DRC that was received the week of March 26, 2007. This returned survey could not be included in the final database due to the late response.

Data Analysis

Dataset Preparation

Of the 1,050 schools and districts that accessed the Website, 1,020 submitted a complete survey and 30 were partially completed surveys. The 30 partial surveys were evaluated for completeness and 14 surveys had enough data to be added to the final dataset (for a total of 1,034 schools and districts). With 1,458 schools and districts invited to complete a survey, 1,034 returned surveys yielded a 70.92% response rate.

After the dataset was finalized, a quality check was conducted to ensure the data fell within appropriate ranges and the comments were cleaned for profanity. Data analysis was conducted on both quantitative data (i.e., answers from the closed-ended questions) and on qualitative data (i.e., the three comment opportunities provided on the survey).

ADJUSTED DATASET

A central purpose of the Web survey was to collect current data from the districts and schools to be used within several components of the Feasibility Study report. Due to the response rate of the survey, a complete set of data including every district and school was not obtained. However, the response rate of 71% did provide enough robustness in the dataset so that it could be used to estimate survey responses for the non-responding districts and schools.

An established technique for dealing with non-response in survey data was used to provide South Carolina with the statewide and district estimates that were necessary for a comprehensive feasibility study. An adjusted data record was prepared for each school that did not respond, and these records provide data that can be used for the Feasibility Study report sections that require a full picture at the state and district levels.

To prepare these records (to compensate for the missing school responses), each district that lacked one or more school responses was examined and the type and number of non-responding schools were identified (e.g., primary, elementary, middle school, or high schools). When a district had responses from 50% or more of the schools of a given type (primary, elementary, middle school, high school), average responses were determined for each type of school based on the responses from the other schools within that district. These data values were entered into a record for each non-responding school. For instance, consider a district in which there are 10 high schools but only eight responded. The records for the remaining two high schools were populated with a series of values that are averages of the responses of the eight high schools that did respond.

When less than 50% of the schools of a given type within a district supplied responses, statewide estimates were used for the schools of that type, within the district.

The adjusted dataset was necessary to provide results for those sections of the feasibility study requiring capacity, capability, and cost estimates. However, the compensated dataset was not appropriate for *all* instances in which survey data are reported. For instance, Components 6–9 required the adjusted dataset because they focus on providing a full picture at both the state and district levels. However, Component 5 did not require this dataset.

Note: Although surveys were submitted by a number of Adult Education Centers and Alternative Schools, these data were not included in the adjusted dataset. Response rates were too low (between about 20% and 30%) to use the averages to adjust the records, and the testing circumstances of these schools appear to vary widely (e.g., survey comments indicated that some adult education centers or alternative schools may not test at all or may test only using facilities of a nearby school).

For this report, the term *original* refers to actual collected data from a district or school and the term *adjusted* refers to actual and adjusted data representing *all* districts and schools (excluding Adult Education Centers and Alternative Schools).

Quantitative Analyses

A variety of quantitative analyses were conducted on the survey data (total state and district levels). This information was supplied to the various writers of the 14 sections of the report where it was needed. For instance, writers of Component 6 on "Hardware, software, staffing, and training requirements at the state, district, and school levels to administer a statewide computer-based or computer-adaptive assessment" made extensive use of information about amount and types of existing technology and people resources to help assess the "gap" between where South Carolina is now and where South Carolina needs to be in order to implement statewide computerized testing.

In addition to the analyses being run specifically for the DRC report writers, DRC has also provided a series of data reports at the district level (showing district and school level data) on the quantitative data.

Qualitative Analyses

There were six (6) open-ended questions on the SC Feasibility Study Web Survey: three (3) at the school level and three (3) at the district level (Table S-2). Open-ended questions provide useful qualitative data that often accompanies quantitative data or closed-ended questions. Comments can be useful in several ways:

1. They provide survey respondents with the opportunity to raise issues that they may see as unique to their situation.
2. They provide a content "safety net" enabling survey respondents to offer input on issues not specifically included in the original survey design (e.g., due to space considerations).
3. They often help to provide more in-depth explanation or illustration of particular concerns revealed by the quantitative results (e.g., quantitative results do very well telling which topics are positive or negative but less well telling why they are positive or negative).

Table S.2: Web Survey Open-ended Questions

School Level	
68.	Thinking about all the needs of your school, please indicate the two or three biggest barriers to implementing computerized testing at your school.
69.	Thinking about all the needs of your school, please discuss the two or three biggest advantages to implementing computerized testing at your school.
70.	Do you have any other comments about how your school could best transition to computerized testing?
District Level	
84.	Thinking about all the needs of your district, please indicate the two or three biggest barriers to implementing computerized testing.
85.	Thinking about all the needs of your district, please discuss the two or three biggest advantages to implementing computerized testing.
86.	Do you have any other comments about how your district could best transition to computerized testing?

There are two sets of reporting deliverables from the qualitative data (comments). One is *verbatim comment reports* and the other is *comment theme report*s. Verbatim comment reports are simply a listing of the verbatim (as typed) comments. Comment theme reports present the major themes (i.e., topics that emerged from the comments themselves). While comment theme reports are derived from the verbatim comments, they do not list every comment. In the comment theme reports the emphasis is on the major themes that emerged from the comments, which are then illustrated with verbatim comments that reflect the theme in question. In order to construct comment theme reports, the comments must first be assigned theme codes. Coding respondent comments increases the usability of the data by allowing for more organization and sorting by theme.

The following steps were taken to create themes for organizing and coding the comments for the comment theme reports. During the administration period of the Web survey, an initial sample of comments (N=79) was downloaded for comment code creation. Next, the sample comments were imported into the SPSS Text Analysis for Surveys 2.0 program. Using both the *Extract Terms* and *Create Categories* features of the software, initial comment themes were created. After an initial review and edit of the themes, all of the comments were imported into the Text Analysis program to further refine the themes. After the DRC and the SC Team reviewed the comment themes, they were finalized. Each comment was assigned up to three comment themes. A final human review also took place to ensure that coding was appropriate and accurate. A list of the final comment themes is provided in Table S-3. Parallel questions 68 and 84, 69 and 85, and 70 and 85 share the same themes.

Table S.3: Comment Themes

Comment Themes for Questions 68 and 84	Comment Themes for Questions 69 and 85	Comment Themes for Questions 70 and 85
Effect on Students Computer skills and abilities, computer access, elementary, high school, student placement	**Effect on Students** Computer skills and abilities, computer access, elementary, high school, student placement	**Effect on Students** Computer skills and abilities, computer access, elementary, high school, student placement
Cost Funding, money, purchase, upgrade, buy, printing	**Cost** Funding, money, purchase, upgrade, buy, printing	**Cost** Funding, money, purchase, upgrade, buy, printing
Time Testing window, testing period, time for preparation, scoring time	**Time** Testing window, testing period, time for preparation, scoring time	**Time** Testing window, testing period, time for preparation, scoring time
Space Location, classrooms, school labs, rooms	**Space** Location, classrooms, school labs, rooms	**Space** Location, classrooms, school labs, rooms
Technology Hardware, computers, Internet access, bandwidth, software, networks	**Technology** Hardware, computers, Internet access, bandwidth, software, networks	**Technology** Hardware, computers, Internet access, bandwidth, software, networks
Test Security Security concerns	**Test Security** Security concerns	**Implementation/Transition** Implementation time
Implementation/Transition Implementation time	**Current Testing Programs** MAPS, PACT	**Current Testing Programs** MAPS, PACT
Special Needs Testing students with special needs	**Electronic Testing (eTesting) Process** Paper, shipping, security, materials, format, packing	**Special Needs** Testing students with special needs
Miscellaneous	**Feedback** Scores, results, turnaround, data	**Miscellaneous**
	Special Needs Testing students with special needs	
	Miscellaneous	

Study on the
Feasibility and Cost of
Converting the
State Assessment Program
to a
Computer-Based or
Computer-Adaptive Format

WEB SURVEY

SURVEY

School-level Survey

- Please enter your school BEDS.

1. Please indicate the title that best fits your role at this school.
   - ○ School technology coordinator
   - ○ School test coordinator
   - ○ School administrator
   - ○ District technology coordinator
   - ○ District test coordinator
   - ○ Other
2. Is your school considered an adult education center? (Yes, No)
3. What is the total number of students in your school for each grade?
   - **a.** Kindergarten
   - **b.** 1
   - **c.** 2
   - **d.** 3
   - **e.** 4
   - **f.** 5
   - **g.** 6
   - **h.** 7
   - **i.** 8
   - **j.** 9
   - **k.** 10
   - **l.** 11
   - **m.** 12
   - **n.** other
4. What is the total number of <u>full-time</u> teachers/instructional staff at your school?
5. What is the total number of <u>part-time</u> teachers/instructional staff at your school?

HARDWARE/SOFTWARE

6. Please indicate the total number of computers available to students at your school.

7. How many classrooms have computers in them for instructional purposes?

8. On average, how many computers are in use for student instruction in each classroom?

As you complete the remainder of the survey, please keep in mind that the questions refer to computers that can be used for testing in a proctored setting. For example, computers that is in computer labs, or classrooms with multiple computers, computers that can be moved/set up in a room for testing, mobile labs, etc.

Computers that are NOT eligible for testing and should NOT be included in your survey answers are: computers in locations that cannot be proctored, computers used for administration purposes, etc.

9. Please indicate the maximum number of computers available (or could be made available) for student testing at your school.

10. My school has enough computers for all students to complete PACT (grades 3–8) testing if the testing window is:

    ○ 1 week

    ○ 2 weeks

    ○ 3 weeks

    ○ 4 weeks

    ○ 5 or more weeks

    ○ My school does not do PACT testing

11. My school has enough computers for all students to complete HSAP (grades 10, 11, 12) testing if the testing window is:

    ○ 1 week

    ○ 2 weeks

    ○ 3 weeks

    ○ 4 weeks

    ○ 5 or more weeks

    ○ My school does not do HSAP testing

12. My school has enough computers for all students to complete EOCEP (grades 7–12) testing if the testing window is:

- ○ 1 week
- ○ 2 weeks
- ○ 3 weeks
- ○ 4 weeks
- ○ 5 or more weeks
- ○ My school does not do EOCEP testing

13. How many classrooms could be used for student testing?

14. How many other rooms could be specifically dedicated to student testing?

15. Please indicate the number of computer labs in your school, including mobile labs.

16. In which area would students most likely be given computerized tests? (Classrooms, Computer labs, Library/Media Centers, Other)

17. For computers that would be used for student testing, please list the number of computers located in each of the following areas:

a. Classrooms: _______________________ computers

b. Computer labs:______________________ computers

c. Library/Media Centers:______________ computers

d. Other: _____________________________computers

18. For computers that would be used for student testing, please list the number of computers:

a. Less than 18 months old: _________ computers

b. Between 19 and 48 months old: ______ computers

c. Older than 48 months: _____________ computers

19. For computers that would be used for student testing, please list the number of computers with the following operating systems:

- **a.** Windows 98: _________ computers
- **b.** Windows NT: _________ computers
- **c.** Windows 2000/ME: ____ computers
- **d.** Windows XP: _________ computers
- **e.** Mac OS 9 or lower: ____ computers
- **f.** Mac OS X: ___________ computers
- **g.** Linux: ______________ computers

20. For computers that would be used for student testing, please list the number of computers with the following amounts of internal memory (RAM):

- **a.** Less than 128 MB: _____ computers
- **b.** 128MB: __________ computers
- **c.** 256 MB: __________ computers
- **d.** 512 MB: __________ computers
- **e.** Greater than 512MB: _____ computers

21. For computers that would be used for student testing, please list the number of computers with a monitor size of :

- **a.** 15" or less: ______ computers
- **b.** 17": ______ computers
- **c.** 19": ______ computers
- **d.** 21": ______ computers
- **e.** 22" or greater: ______ computers
- **f.** Other: ______ computers

22. For computers that would be used for student testing, please list the number of computers with a monitor resolution of:

- **a.** Less than 800x600: ______ computers
- **b.** 800x600: ______ computers
- **c.** 1280x720: ______ computers
- **d.** 1280x1024: ______ computers
- **e.** 1440x900: ______ computers
- **f.** 1600x1200 or greater: ______ computers

23. For computers that would be used for student testing, please list the number of computers with the following processor speeds:

a. Less than 200 MHz: ______ computers

b. 200 MHz to 500 MHz: ______ computers

c. 500 MHz to 1GHz: ______ computers

d. Greater than 1GHz: ______ computers

24. For computers that would be used for student testing, do they all have access to the Internet? (Yes/No)

25. If no, what number of computers that would be used for student testing do not have access to the Internet?

26. Approximately how would you describe overall Internet usage at your school at different times of the day?

	Heavy	**Moderate**	**Light**	**None**
a. Before school				
b. During morning classes				
c. During lunch time				
d. During afternoon classes				
e. After school				

27. For computers that would be used for student testing, please estimate the number of computers connected to the Internet through the following means:

a. Dial-Up/Modem: ______ computers

b. Cable Modem or DSL: ______ computers

c. T1: ______ computers

d. T3 or greater: ______ computers

e. Other: ______ computers

28. Of the computers that would be used for student testing, please estimate the number that use the following browsers for Internet access:

- **a.** Internet Explorer: ______ computers
- **b.** Netscape Navigator: ______ computers
- **c.** Apple Safari: ______ computers
- **d.** Mozilla: ______ computers
- **e.** Firefox: ______ computers
- **f.** Opera: ______ computers
- **g.** Other: ______ computers

29. For computers that would be used for student testing, does your school use any type of 3rd party software to track Internet usage? (Yes, No, Don't know)

30. For computers that would be used for student testing, does your school use any type of 3rd party software to block Internet content? (Yes, No, Don't know)

31. Do you allow the computers that would be used for student testing to connect out to the Internet using HTTP and HTTPS protocols (port 80 and 443)? (Yes, No, Don't know)

32. For computers that would be used for student testing, does your school use any type of firewall product? (Yes, No, Don't know)

33. If yes, indicate if the firewall product is one of the following:

- ○ Outpost Firewall Pro
- ○ Check Point VPN-1 (formerly Firewall-1)
- ○ Kerio WinRoute Firewall
- ○ Microsoft Internet Security and Acceleration Server
- ○ Symantec
- ○ Cisco Pix
- ○ Sonic Wall
- ○ Other

34. Select the connectivity type that best describes your location (Traditional CATx, Wireless, Other)

35. Do any of the computers that would be used for student testing use wireless networking for Internet access? (Yes, No, Don't know)

36. If yes, for computers that would be used for student testing, please indicate the approximate percentage of total computers in your school that utilize wireless networking.

37. Is your wireless network using any form of encryption? (Yes, No, Don't know)

38. For computers that would be used for student testing, are the hard drives of these computers backed up on a regular basis? (Yes, No, Don't know)

39. If yes, how frequently are they backed up?

- More than once a day
- Once a day
- Once every 2 days
- Once every 3 days
- Once a week
- Less than once a week

40. Are the backup media stored offsite? (Yes, No)

41. Please indicate any specialized software/hardware available in your school to assist special needs students *(Mark all that apply)*

a. ZoomText

b. TextReader

c. Large print keyboard

d. Dictaphones

e. Braille Blazer

f. Speaking dictionary

g. Expert mouse pro, for easy cursor manipulation

h. Hear-it amplifying devices

i. Headsets or headphones

j. Braille Translator

k. Other

42. My school has an adequate budget to maintain computers that would be used for student testing.

- Strongly agree
- Agree
- Neither agree nor disagree
- Disagree
- Strongly disagree

43. My school would require additional funding to acquire the computers necessary for student testing.

- Strongly agree
- Agree
- Neither agree nor disagree
- Disagree
- Strongly disagree

44. Do you have plans to upgrade your student computer fleet? (Yes, No)

45. If yes, when is the upgrade planned to take place?

- Less than 6 months
- 6 months to 1 year
- In about 2 years
- In about 3 years
- In about 4 years
- In about 5 years or more

46. On average, about how much is currently spent each year to maintain your school computers (e.g., average over the last 5 years)?

47. On average, about how much is currently spent each year to upgrade/buy new computers (e.g., average over the last 5 years)?

48. What percentage of your school budget is dedicated to computer purchases and computer maintenance for student computers?

49. On average, what percentage of student computers in your school is replaced each year?

STAFFING/TRAINING

50. Please indicate the total number of Information Technology (IT) staff available at your school.

51. What types of IT personnel are available for your school? *(mark all that apply)*

- **a.** Network Administrator
- **b.** Database Administrator
- **c.** Technology Director
- **d.** IT Director
- **e.** Information Systems Manager
- **f.** Media Specialist
- **g.** District-based technician(s)
- **h.** Contracted (outsourced) vendor support
- **i.** No school-based support provided
- **j.** Other

52. Please indicate the number of temporary personnel you needed during the 2005–2006 test administrations (e.g., monitors, volunteers, temps, etc).

53. Please indicate the number of teachers you currently train annually for test administration.

54. Please indicate the number of school technical coordinators you currently train annually for test administration.

55. Please indicate the number of other test administrators (excluding teachers, technical coordinators, and temporary staff) you currently train annually for test administration.

56. Is your current school involved in computerized testing? (Yes, No)

57. What is your level of experience with computerized testing?

- ○ I have never been involved at all with computerized testing.
- ○ I have been involved one or two (1 or 2) times with computerized testing.
- ○ I have been involved three to five (3 to 5) times with computerized testing.
- ○ I have been involved more than five (> 5) times with computerized testing.

58. Please estimate the number of temporary personnel you would need during test administration for a computerized test.

59. Please estimate the number of teachers you would need to train for test administration for a computerized test.

60. Please estimate the number of school technical coordinators you would need to train for test administration for a computerized test.

61. Please estimate the number of other test administrators (excluding teachers, technical coordinators, and temporary staff) you would need to train for test administration for a computerized test.

STUDENT FAMILIARITY WITH COMPUTERS

62. At what grade levels are computers regularly used for classroom instruction at your school?

a. Kindergarten
b. 1
c. 2
d. 3
e. 4
f. 5
g. 6
h. 7
i. 8
j. 9
k. 10
l. 11
m. 12

63. How frequently do students at your school practice writing or composing on computers?

○ Daily, 3–4 times a week, 1–2 times a week, Less than once a week, Less than once a month, Never, Don't know.

64. At each grade level at your school, please estimate the amount of time students spend on a computer during a typical school day.

	None	0–30 minutes	30–60 minutes	60–90 minutes	More than 90 minutes
a. Kindergarten					
b. Grade 1					
c. Grade 2					
d. Grade 3					
e. Grade 4					
f. Grade 5					
g. Grade 6					
h. Grade 7					
i. Grade 8					
j. Grade 9					
k. Grade 10					
l. Grade 11					
m. Grade 12					

65. At your school, please indicate the lowest grade level at which students receive direct or formal instruction in keyboarding skills.

66. What type(s) of word-processing software are used by students in their writing/composing on the computer? *(mark all that apply)*

a. AppleWorks

b. Mariner Write

c. Mellel

d. Microsoft Word

e. Microsoft Works

f. TextEdit

g. WordPerfect

h. Other

67. What tools are available to the students on the computers? *(mark all that apply)*

- **a.** Spell-checker
- **b.** Grammar-checker
- **c.** Dictionary
- **d.** Thesaurus
- **e.** Calculator
- **f.** Graphing calculator
- **g.** No tools are available/enabled
- **h.** Other

Open-ended Questions

68. Thinking about all the needs of your school, please indicate the two or three biggest barriers to implementing computerized testing at your school.

69. Thinking about all the needs of your school, please discuss the two or three biggest advantages to implementing computerized testing at your school.

70. Do you have any other comments about how your school could best transition to computerized testing?

District Survey

- Please enter your district BEDS.

71. Please indicate the title that best fits your role at this district.

- ○ District technology coordinator
- ○ District test coordinator
- ○ Other

72. What is <u>your level</u> of experience with computerized testing?

- ○ I have never been involved at all with computerized testing.
- ○ I have been involved one or two (1 or 2) times with computerized testing.
- ○ I have been involved three to five (3 to 5) times with computerized testing.
- ○ I have been involved <u>more than</u> five (> 5) times with computerized testing.

HARDWARE/SOFTWARE

73. Do your district servers use any type of firewall product? (Yes, No, Don't know)

74. If yes, indicate if the firewall product is one of the following:

- ○ Outpost Firewall Pro
- ○ Check Point VPN-1 (formerly Firewall-1)
- ○ Kerio WinRoute Firewall
- ○ Microsoft Internet Security and Acceleration Server
- ○ Symantec
- ○ Cisco Pix
- ○ Sonic Wall
- ○ Other

75. Are the hard drives of the district servers backed up on a regular basis? (Yes, No, Don't know)

76. If yes, how frequently are the servers backed up?

- ○ More than once a day
- ○ Once a day
- ○ Once every 2 days
- ○ Once every 3 days
- ○ Once a week
- ○ Less than once a week

77. Are the backup media stored offsite? (Yes, No)

78. My district has an adequate budget to maintain computers that would be used for student testing.

- ○ Strongly agree
- ○ Agree
- ○ Neither agree nor disagree
- ○ Disagree
- ○ Strongly disagree

79. My district would require additional funding to acquire the computers necessary for student testing

- ○ Strongly agree
- ○ Agree
- ○ Neither agree nor disagree
- ○ Disagree
- ○ Strongly disagree

80. Does your district have plans to upgrade your student computer fleet? (Yes, No)

81. If yes, when is the upgrade planned to take place?

- ○ Less than 6 months
- ○ 6 months to 1 year
- ○ In about 2 years
- ○ In about 3 years
- ○ In about 4 years
- ○ In about 5 years or more

STAFFING/TRAINING

82. Please indicate the number of Information Technology (IT) staff available at your district level.

83. What types of IT personnel are available at your district level? *(mark all that apply)*

- **a.** Network Administrator
- **b.** Database Administrator
- **c.** Technology Director
- **d.** IT Director
- **e.** Information Systems Manager
- **f.** Media Specialist
- **g.** District-based technician(s)
- **h.** Contracted (outsourced) vendor support
- **i.** Other

OPEN-ENDED QUESTIONS

84. Thinking about all the needs of your district, please indicate the two or three biggest barriers to implementing computerized testing.

85. Thinking about all the needs of your district, please discuss the two or three biggest advantages to implementing computerized testing.

86. Do you have any other comments about how your district could best transition to computerized testing?

EXPERT PANEL MEETING OFFICIAL MINUTES

Expert Panel/Advisory Committee Panel Meeting

Wednesday, March 28, 2007

9:00 AM – 3:00 PM

Inn at USC – Columbia, SC

Carolina Room

Minutes:

9:00 – Tammy Mainwaring, Project Manager, welcomed all guests and turned the meeting over to Pat Porter – DRC Vice President of Large Scale Assessment

Ms. Porter talked about the purpose of the meeting in tapping into the knowledge and expertise of the expert panel. She reviewed the 14 components contained in Act 254. She also introduced the moderator – Dr. Steven Wise – Moderator - James Madison University, and the expert panel member.

Dr. Steven Wise:

Reviewed agenda topics (attached):

Each panel member introduced themselves and described their state roles and gave a brief overview of computer-based or computer-adaptive testing in the following order:

Karen Echeverria – Idaho

Mildred Bazemore – North Carolina

Sarah Susbury – Virginia

Kristin Ellington – Florida

Wes Bruce – Indiana

Jennifer Stegman – Oklahoma

Sandra Foster – West Virginia

Roger Ervin – Kentucky

John Poggio – University of Kansas, Computerized Assessments and Learning

After introductions and the high level overview, Dr. Steven Wise asked for clarification questions.

Question – What makes voluntary successful or not successful?

Answer –

> Indiana: One issue is matching calendars with school districts. Some states had more leverage with schools than others.
>
> Florida – Layering on an environment on top of something that is already working for them is a concern. Cheating is a concern. Technology support is a concern. The technology gap and infrastructure problems are a concern.
>
> Oklahoma – The voluntary study was extra work on districts. The key is to make sure you count student scores. You should make sure the districts and schools get something out of the voluntary study.

Questions - What about technology gaps and inadequate infrastructure?

Answer –

North Carolina – conducts a feasibility study that addresses challenges, connectivity, and bandwidth at the school level. North Carolina has mandated several tests to be delivered online. In places where there are deficits, the states provide resources such as mobile labs to help with problems.

Indiana – the state should review Virginia's implementation plan. In retrospect, Indiana would have done this. One lesson learned is to store nothing in the RAM of the computer. Indiana captures information after each item so work is not lost.

You have to also use computers for instruction as well.

Kentucky – RFP – you must decide what software/hardware you are going to support. A state cannot support everything. Things such as security patches sometimes cause technology not to work the next day. States should incorporate test plans and back up plans.

Question – Advice on implementation?

Answer-

John Poggio – University of Kansas – Implementation – A state needs experience and expertise – it is needed in technology and testing and education. All teams should have these three audiences at the table. Kansas started small with multiple choice and certain grade levels. Kansas began with a volunteer program. The state should work with universities to implement programs.

Question: When you implemented computerized testing, was the goal to raise student achievement or to reduce testing cost? Can you tell me about the impact of computerized testing on student achievement?

Answer –

Florida – They want to provide world class education. Testing needs to be relevant to the workplace students will enter.

Virginia – They need scores by quickly for end-of-course tests – especially for students who retake. The initial investment for computerized testing was not a cost savings.

North Carolina – The state wanted to enhance the testing experience for students and make it more authentic.

Question: The districts have varying degrees of capacity to support computerized testing. Therefore the districts with better financial means can participate. What are your states doing to help less wealthy districts participate?

Answer –

Idaho – More state technology funding went to districts which needed technology to implement computerized testing.

Virginia – Funding based on number of schools in district. Virginia's technology had been emphasized before this initiative so no districts were far behind. They made sure that all rural districts were trained.

Florida – Rural schools are Title 1 schools and could put that money toward testing. It must be a leadership priority.

Oklahoma – It is not the hardware that is the problem but having people with expertise and technical skills to implement computerized testing.

Question – How comparable are online testing scores with paper/pencil scores?

Answer -

Idaho – saw an increase in the scores with online testing – hypothesis – kids are used to dealing with computers.

Indiana – They had comparability for the overall scores but not on subtests – such as graphing plots on a graph. The state should look at possible mismatch of instruction and how they are testing. Comparability in English is not as much of an issue as Algebra II. It now appears that students are doing betting in the online version of testing in English.

Virginia – Critical point – make sure instruction matches testing. Over time, comparability has become more prevalent.

North Carolina – They are now looking at end-of-course test performance comparability. This state requires all students to test online [for Physics test only].

Richard Luecht – Univesity of North Carolina-Greensboro (UNC-G) discussed research studies addressing comparability of scores around several factors.

John Poggio – Comparability of computerized testing and paper/pencil testing is a short lived phenomenon – until the computer completely penetrates all of society.

Discussion:

Indiana - One grade level of paper/pencil testing in Indiana takes three tractor trailer loads of paper answers to be delivered to be graded….this goes away when you go online.

Kentucky – Testing industry needs to generate a "standard toolkit" such as equations editors for algebra should be provided for schools to use during the year and on the test.

Oklahoma – issues a practice test so students can adapt to online format and tools.

Kansas – tools are released as public domain so students can use at schools, home, etc. Kansas develops test tutorials, item tutorials, etc. The test should support instruction.

West Virginia – In writing there is a natural progression from working in a lab setting to write essays and moving to computerized testing. The formative assessment should also be put in RFP along with summative.

John Poggio, Kansas – If every state could take one component of computerized testing and really do it well – by the end of the day, you would have 50 components done extremely well – states should share knowledge and best practices – there should be a national consortium to address these issues.

Kentucky – As we get closer to 2014, (NCLB) – leaders in schools are hesitant to change – computerized testing is a huge change.

Question (SC Policy Council) - – How many states here use DRC?

Answer -

Idaho - Yes

North Carolina – No

Virginia – No

Florida – No

Indiana – No

Oklahoma – Yes

West Virginia – No

Kentucky – No

Kansas - No

Question: Will there be any states here that do not use computerized testing that will be represented as the study calls for it?

Answer -

Project design is to include surveys for all states and the information will be gleaned from states that have chosen not to implement computerized testing. More mileage can be gotten from a focus group of states which have experienced some form of computerized testing to talk about lessons learned, advantages, disadvantages, etc.

Question: How can computerized testing inform state policy? Can this be done for all students?

Answer -

Kansas – This can be done for all students. Instruction needs to match testing. 98% of students prefer taking tests online. Computerized testing is the future – you will see it come. Families will expect this in the future. Society will not turn away from computers – look at the way industry is moving.

Virginia – Think about certification exams that students will have to take for the workplace – there is more testing online in industry now than in education. Education is catching up with industry.

Richard Luecht (UNC-Greensboro) – State must be clear about purpose. For example, diagnostic tests is a different cost than other forms of testing. It is how much value you put in a measurement.

Question - Please describe testing windows.

Answer -

Indiana – the state has not gone online with accountability tests yet because they do not have an adequate amount of computers to accomplish this. For tests that are offered online, there is a 2 week testing window. $100.00 laptops already exist – this will be a reality with all tests one day.

North Carolina – Access, time scheduling, and policies have dictated testing windows. End-of-course tests are optional as to whether students want to take online or on paper – because of concerns about security, access, etc. North Carolina also stated that it did not have enough computers to test all students on all tests.

Richard Luecht (UNC Greensboro) - Biggest cost drivers in CBT are seat time and item production.

Question: Are any states conducting Web based testing to mitigate capacity demands?

Answer -

Kentucky - Risks in breakage for secure assessment are too great. They are moving to the local level for capacity and storage.

Virginia – Material remains encrypted – districts are required to use caching. Virginia is using contractors to host.

Question: Are open response items cheaper than closed response?

Answer –

Richard Luecht (UNC Greensboro) – It is more expensive because rubrics must be created to grade these. There are no particular cost savings per item type.

Security Discussion:

Virginia - Elementary school setups of 6 computers clustered together in classrooms is not a good set up for online testing.

Indiana – a proctoring system is essential to ensure that students are taking the test and doing the work.

Oklahoma - When you use Macs and PCs – security requirements will be different. Two security plans must be in place.

Richard Luecht (UNC Greensboro) – Technical experts must be at the table to ensure all technical questions and concerns are addressed.

Virginia – They want a standardized environment where all students have the same amount of real estate on the page. Computers could be used for instruction as well – palm pilots could not. Local districts should be involved and in communication loop regarding requirements for testing. Something should be built into the system to save responses so you don't lose them if the system goes down.

North Carolina – Annual evaluations make sure all upgrades and systems are compatible. Advised to plan ahead and test to make sure all systems are working and compatible.

Florida – The states and NCLB are pushing testing industry with requirements. Three RFP's are currently out and these states are concerned that they will not have bidders.

Discussion: Staffing and Training

West Virginia – Instruction will drive success in computerized testing. Teachers must be comfortable using computers in instruction. They should be provided professional development to do so.

Steven Wise, James Madison University – Instructional people must talk to IT people to make sure standards are the same.

Virginia – Two elements of training – adult training (administrative) and student training (how do I navigate through the test and use the testing tools) – using Web ex and video training modules for adult training.

Kentucky – training with students – a practice test is invaluable. With adults, Webinars and technology training.

Kansas – Before the test, build online tutorials for teachers, students, proctors, etc. Practice tests are a part of this. Also use web conferences, video teleconferencing, and help desk services. Weekly updates are provided to administrators, teachers, and technology personnel with FAQs, etc. They tell the schools that for every 10 days of testing there will be 1 bad day (just like a snow day). Make sure you have a help desk number available.

Indiana – You must have a help desk and the help desk must be very responsive – states must plan to have people to go out "live" to districts when needed. Redundancy is important. Intense monitoring of district [server] farms is required. They need to have the ability to redirect.

Kentucky – lesson learned – do not test on Monday morning. There is a lot of financial traffic that will slow down the network.

North Carolina – Can you recover and at what point should be asked in RFP.

Question – Were invitations to participate in computerized testing issued to districts of all sizes, SES status, rural, urban, etc.?

Indiana – You should actively recruit districts at all levels. State must recruit to make sure that all areas are covered. You must find a way to get a good representation of the state.

North Carolina – Offers incentives for districts to participate. Allowed students to keep the higher of the two scores.

Kansas – Built networks within communities to share success stories for computerized testing. Teachers should be kept informed and their buy-in is necessary.

Discussion: Greatest Challenges for Special Populations

Kentucky: Lack of familiarity of vendors with the special population and their learning needs.

Florida – Uses text readers

Questions – What are the hidden costs in security in computerized testing?

Answer –

Virginia – Privacy filters are very expensive – some use cardboard dividers – they use computer towers to block screens, etc.

Kansas – Use form codes and makes sure same form codes are not sitting beside one another.

Advice to SC from Expert Panel Members:

Idaho – Administers 2nd through 10th (LA, Reading, Math, History and Science). Graduation requirement is online. SC needs to set expectations and goals of computerized testing. Ensure that we have technology, bandwidth, and hardware to successfully implement computerized testing. Make sure RFP is very specific to requirements. Make sure legislature is ready to fund up front costs. Training for staff and students is critical. State representatives must be at training to answer policy questions. Be prepared for resistance. Idaho would implement computerized testing again if given the opportunity.

North Carolina – Administers 10 end-of-course online tests as an option – computer skills test is online - NC would implement computerized testing again if given the opportunity - absolutely. Plan well – lay out an implementation plan involved all stakeholders. Phase in the tests we offer online. Monitor your bandwidth and capacity at the building level.

Virginia – Administers 11 end of course tests online. Also grades 3-8 at every level and science in grades 5 and 8. They are all part of state and federal accountability program. Virginia would implement computerized testing again if given the opportunity. SC should know purpose for embarking on computerized testing. Construction of RFP and evaluation process are critical. Make sure education, assessment and technology experts are at the table. Network with other states and share resources. Remember students are stakeholders too. Project management is critical.

Florida – Administers graduation retest exams in reading and math via computer three times each year. Also computer literacy assessment is given online. Make sure you have an advisory committee for your computerized testing initiative. Use scientifically based research to guide your decisions. Include security audit of vendor software in RFP. Florida would implement computerized testing again if given the opportunity.

Indiana – Administers end course tests online**.** Don't port paper test over to online environment – build your test for online. Look for advice and look at what other states have done.

Oklahoma – Administers geography, math and reading for 7th grade online. Optional biology and history online. Pilot tests are critical to inform decisions. Online testing isn't really less work than paper based, just different work with different groups of people.

West Virginia – Administers 7th and 10th grade writing assessments. NAEP is going online with the writing test in 2011. In RFP, tie in practice tests and formative assessments. West Virginia would implement computerized testing again if given the opportunity.

Kentucky – Administers online tests special populations, reading, math 3-8, science – grade span –need support of Superintendent of Education and State Board. Make sure you plan for adequate technology support

Study on the Feasibility and Cost of Converting the State Assessment Program to a Computer-Based or Computer-Adaptive Format

State
Data Survey Results Report

June 2007

Table of Contents

Introduction

Background

The South Carolina General Assembly passed legislation (Act 254 of 2006) to implement the recommendations of the 2004-2005 Testing Task Force that called for a study of computer-based or computer-adaptive formats for the statewide testing program. The State Budget and Control Board (B&CB), with the advice of a group convened by the Education Oversight Committee (EOC) and State Department of Education (SDE), contracted with Data Recognition Corporation (DRC) to conduct this study.

As part of this study, DRC asked District Test Coordinators and school administrators to complete a survey about current computer and network capabilities at both district and school levels. This input provided critical information that will help the State better understand the technology capacity and staffing available for a computer-based or computer-adaptive testing program. This report includes a summary of the responses to the district and school survey questions for the State.

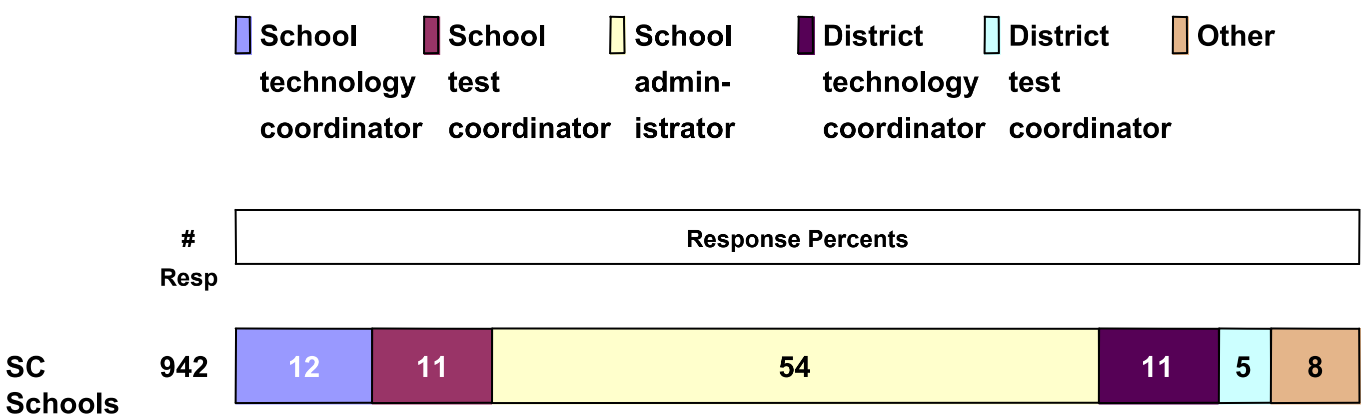
How to Use This Report

Report Layout

The results from the survey are divided into two sections: District Survey Results and School Survey Results. The District Survey Results section contains a summary of responses from all South Carolina school districts that responded to the District Survey questions. The School Survey Results section contains summarized responses from the School Survey questions for all responding schools across the State.

Survey results for most questions are displayed in a bar graph format that shows percentages of responses for each answer choice. The aggregate of all responses to the particular question may not sum to exactly 100% due to rounding. An example of this format appears below. The number of respondents (# Resp) notes the number of districts or schools that provided a response to that particular survey question.

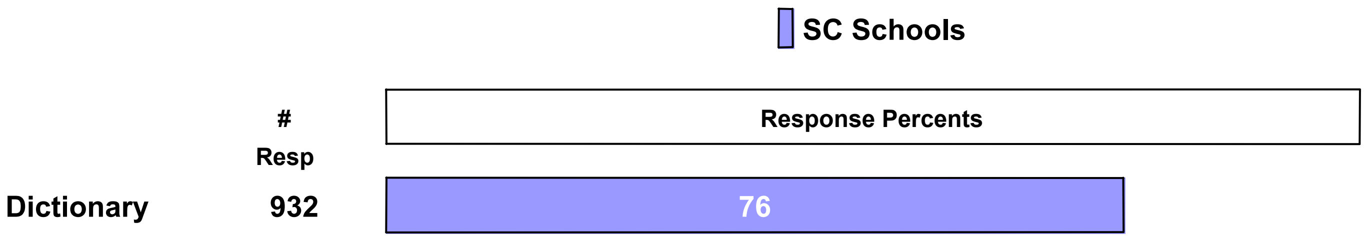
Please indicate the title that best fits your role at this school.



How to Use This Report

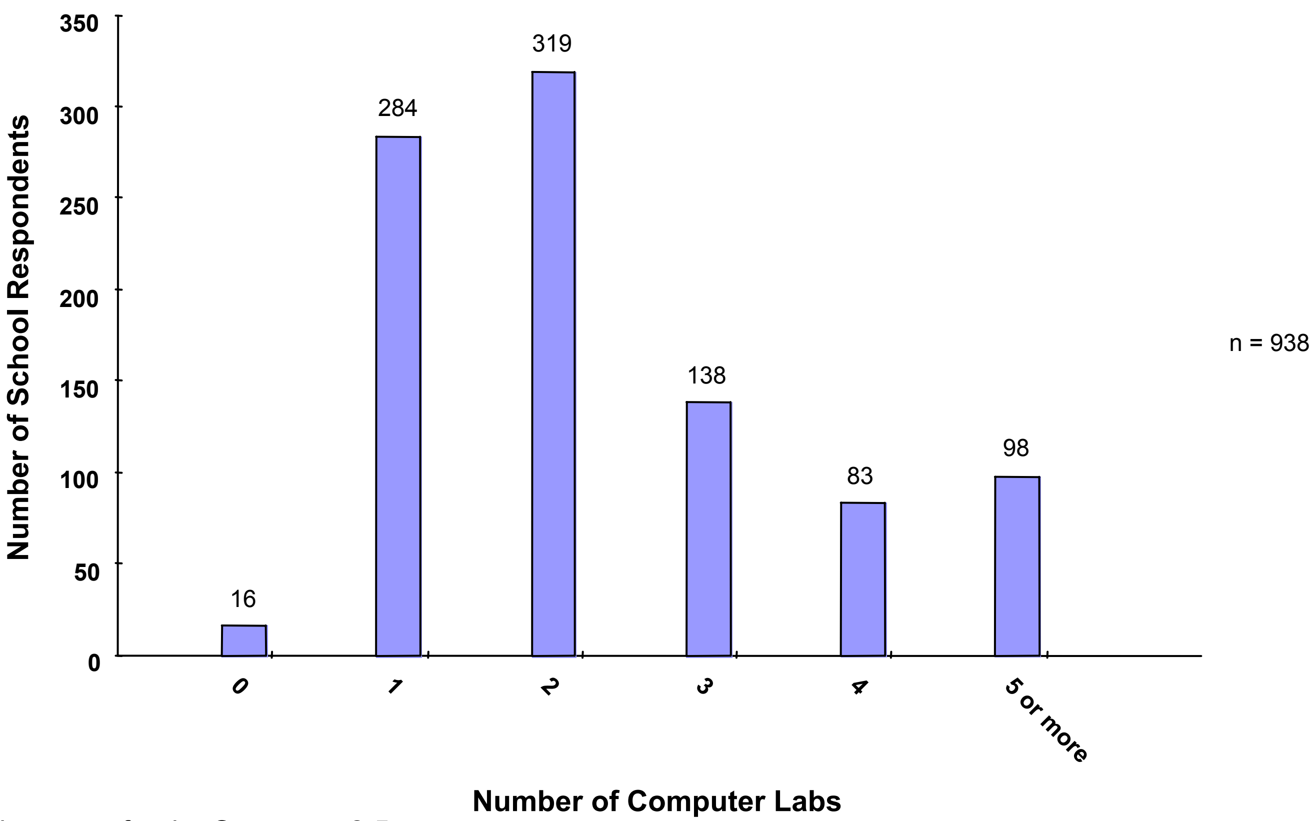
The results of some survey questions are presented in a standard horizontal bar graph. The length of the bar indicates the percentage of responses to the specified answer option. An example of this format appears below.

What tools are available to the students on the computers?



Finally, other survey results are presented in a vertical bar graph. The height of each bar represents the number of survey respondents who provided a response for the stated question. The total number of individuals responding to the question is noted by “n=“. Where appropriate, the mean for the State will be noted below the graph. An example of this format appears below.

Please indicate the number of computer labs in your school, including mobile labs.



The mean for the State was 2.5.

How to Use This Report

Report Terminology

Throughout the report, various terms are used in the presentation of the data. The following is a complete listing of the different terms used throughout the report:

SC Districts – percents of responding districts selecting the noted responses.

SC Schools – percents of responding schools selecting the noted responses.

 # Resp (Number of Responses) – identifies the number of districts or schools responding to the question. This number may vary from question to question.

% Positive – a sum of the percentages for the two positive responses (Strongly Agree and Agree) on a Strongly Agree to Strongly Disagree response scale.

Mean – average response for a question calculated at the State level.

Mean No Internet Access – average response for the number of computers that may be used for student testing but are not connected to the Internet.

 # Students – the number of enrolled students that survey respondents reported for their school. An aggregation of these numbers was used to report the total numbers of students for the State.

 # Teachers – the number of teachers that survey respondents reported for their school. An aggregation of these numbers was used to report the total number of teachers for the State.

 # Rooms – the number of rooms (e.g., classrooms and other rooms) that survey respondents reported for their school. An aggregation of these numbers was used to report the total number of rooms for the State.

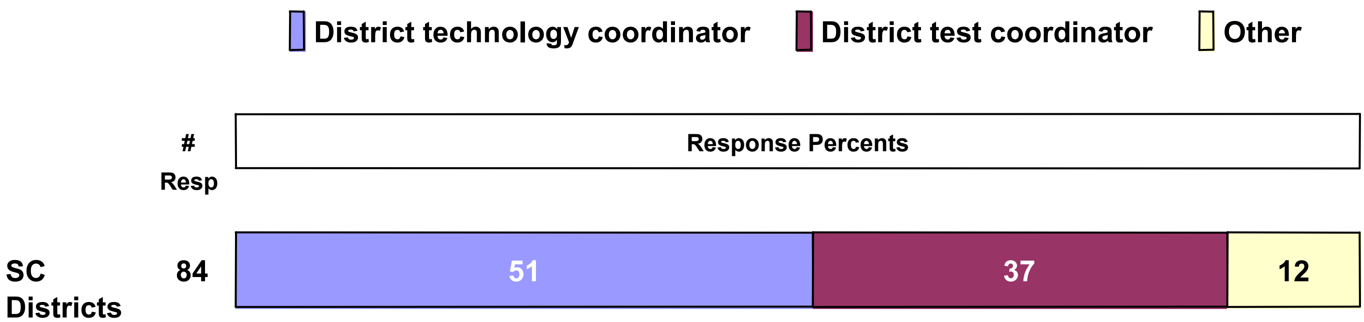
 # Computers – the number of computers that survey respondents reported for their school. An aggregation of these numbers was used to report the total number of computers for the State.

 # Staff – the number of test administration staff (e.g., teachers, school technical coordinators, and test administrators) that survey respondents reported for their school. An aggregation of these numbers was used to report the total number of staff for the State.

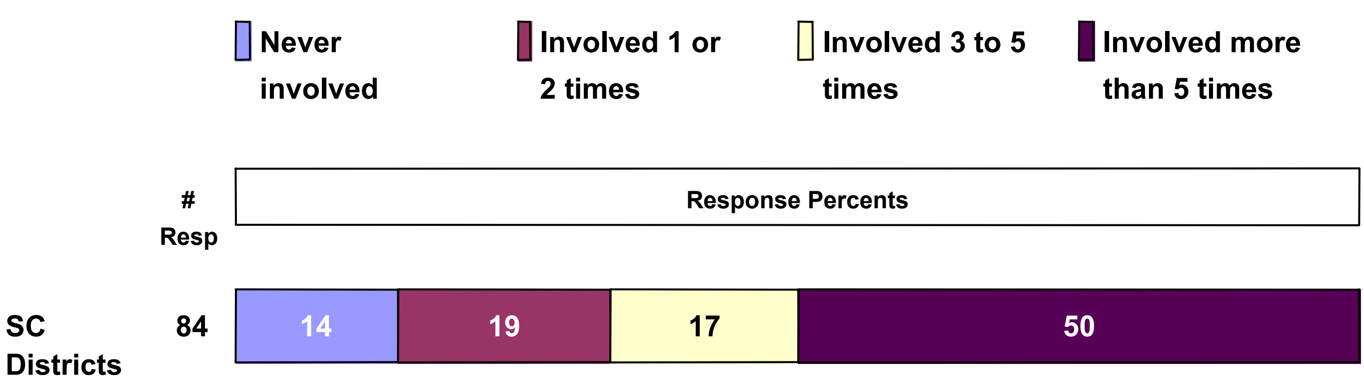
District Survey Results

District Demographics

Please indicate the title that best fits your role at this district.

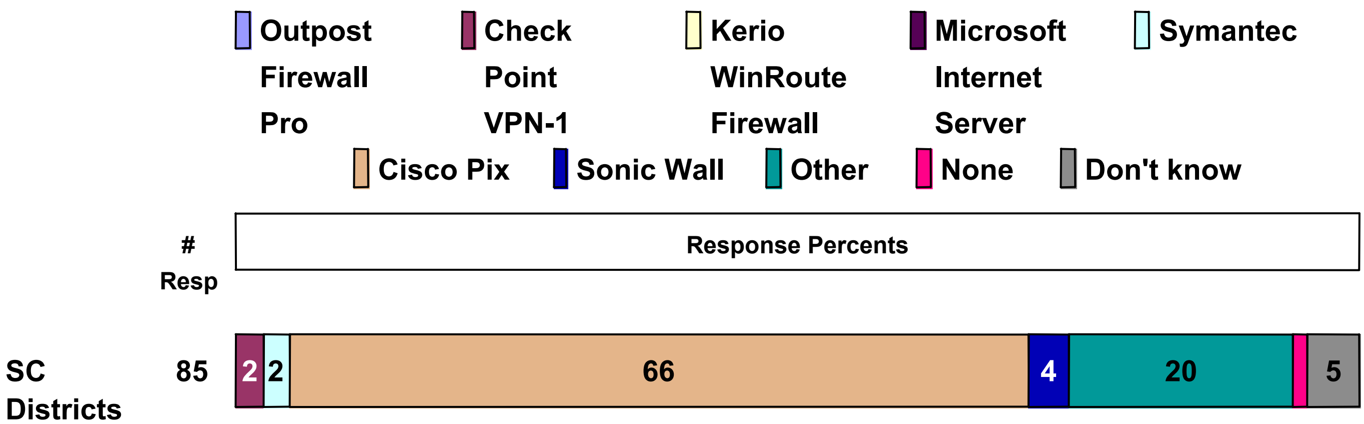


What is your level of experience with computerized testing?

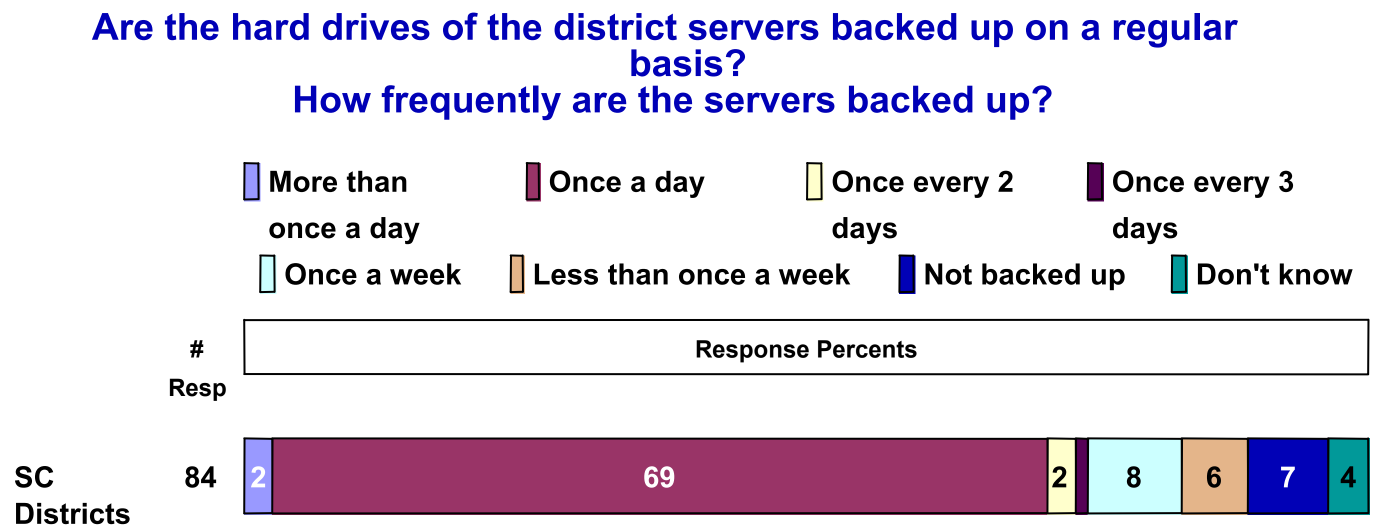


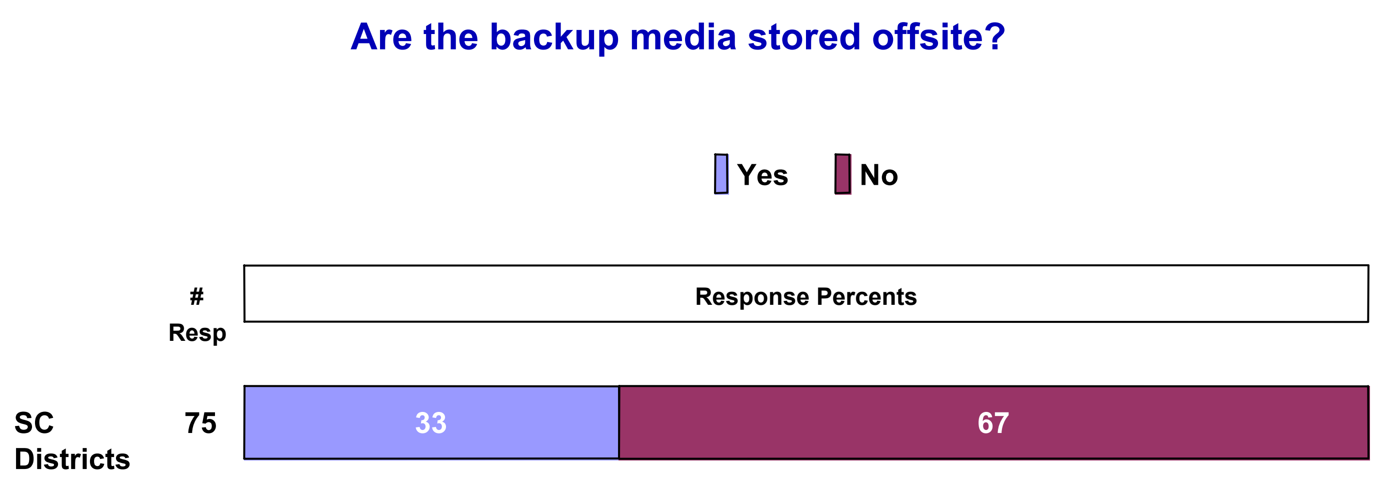
District Firewalls

Do your district servers use any type of firewall product? Indicate if the firewall product is one of the following.



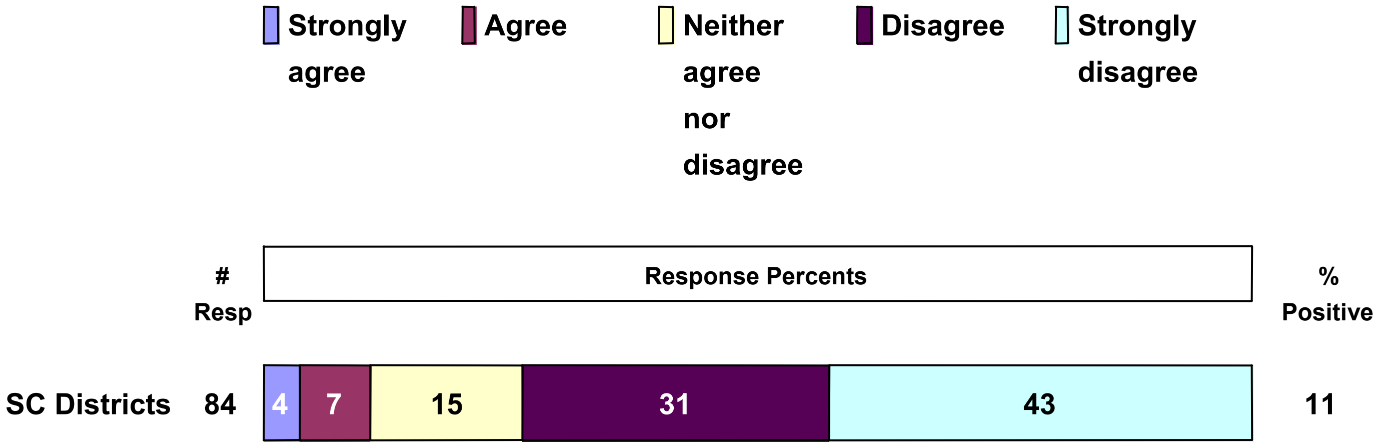
District Backup/Storage



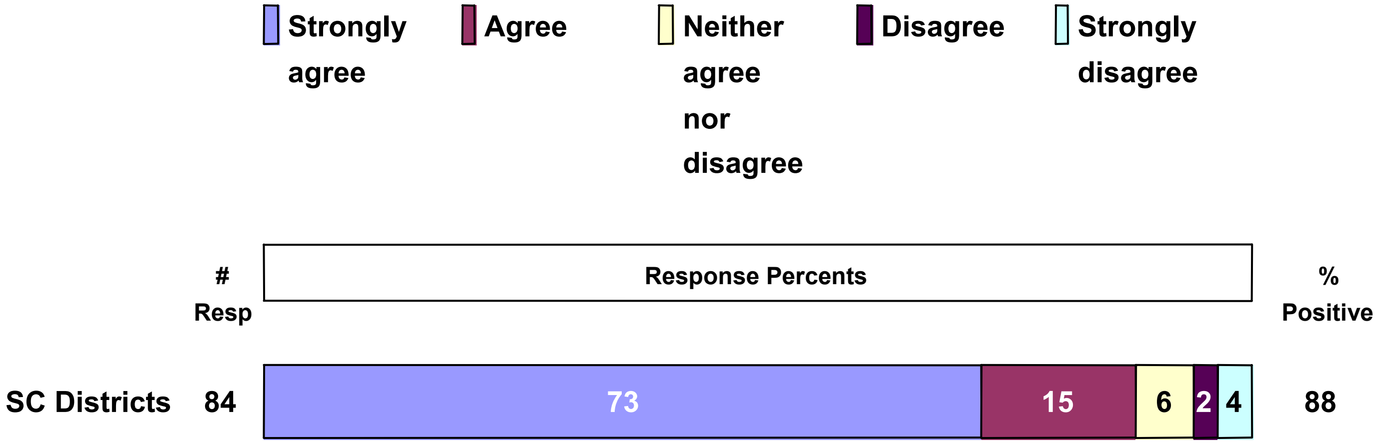


District Refresh Budgets

My district has an adequate budget to maintain computers that would be used for student testing.

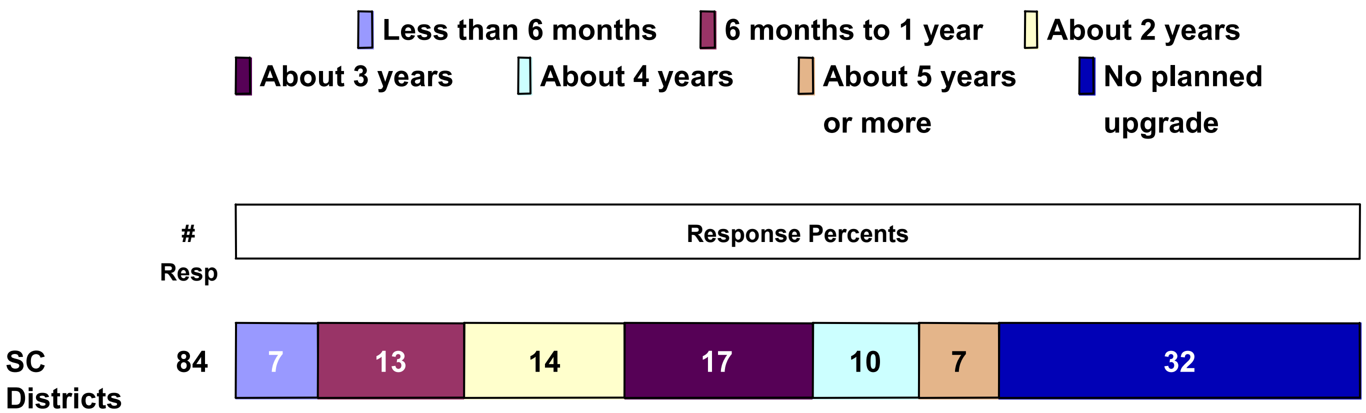


My district would require additional funding to acquire the computers necessary for student testing.



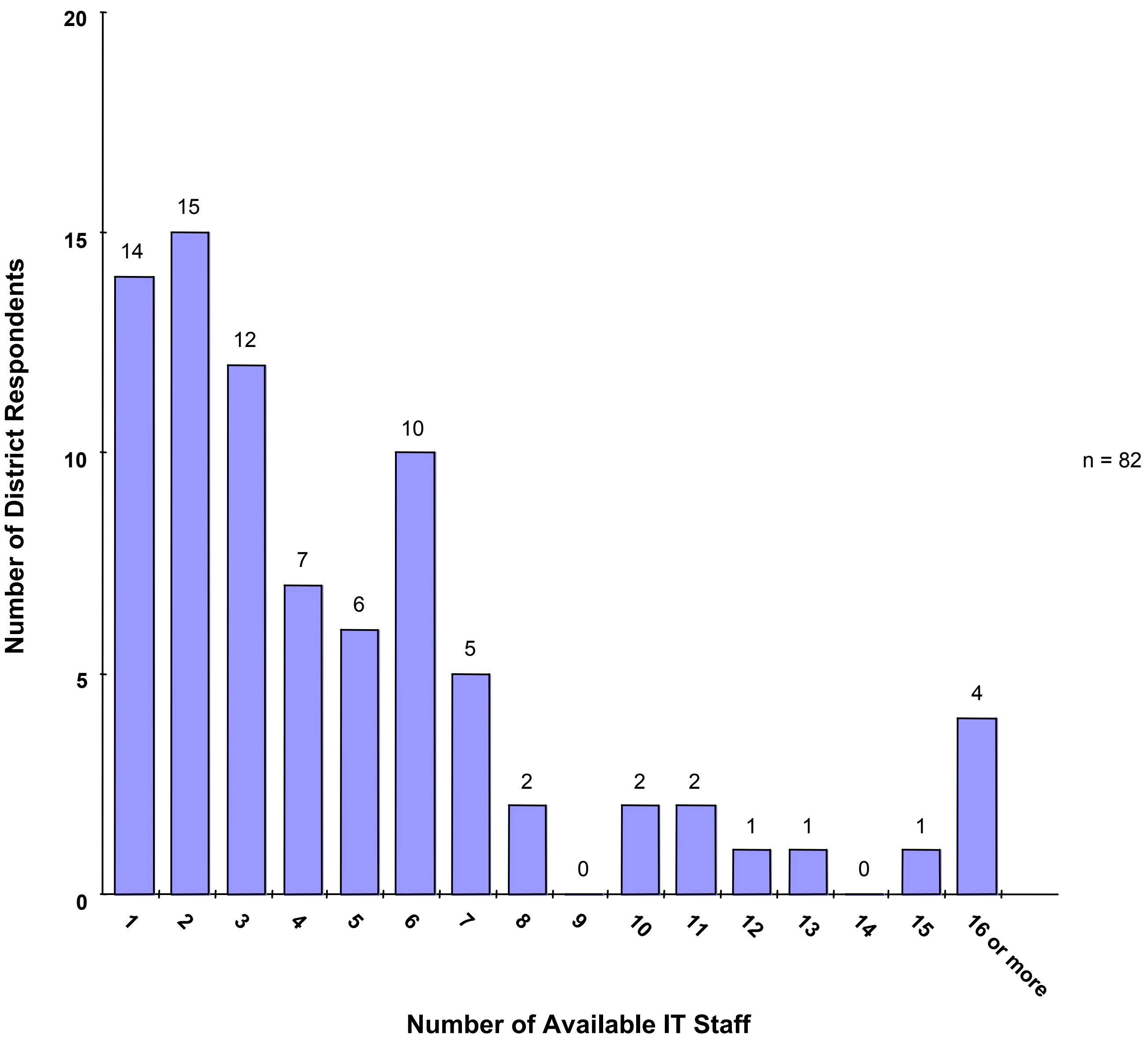
District Refresh Budgets

Does your district have plans to upgrade your student computer fleet? When is the upgrade planned to take place?



District Staffing/Training

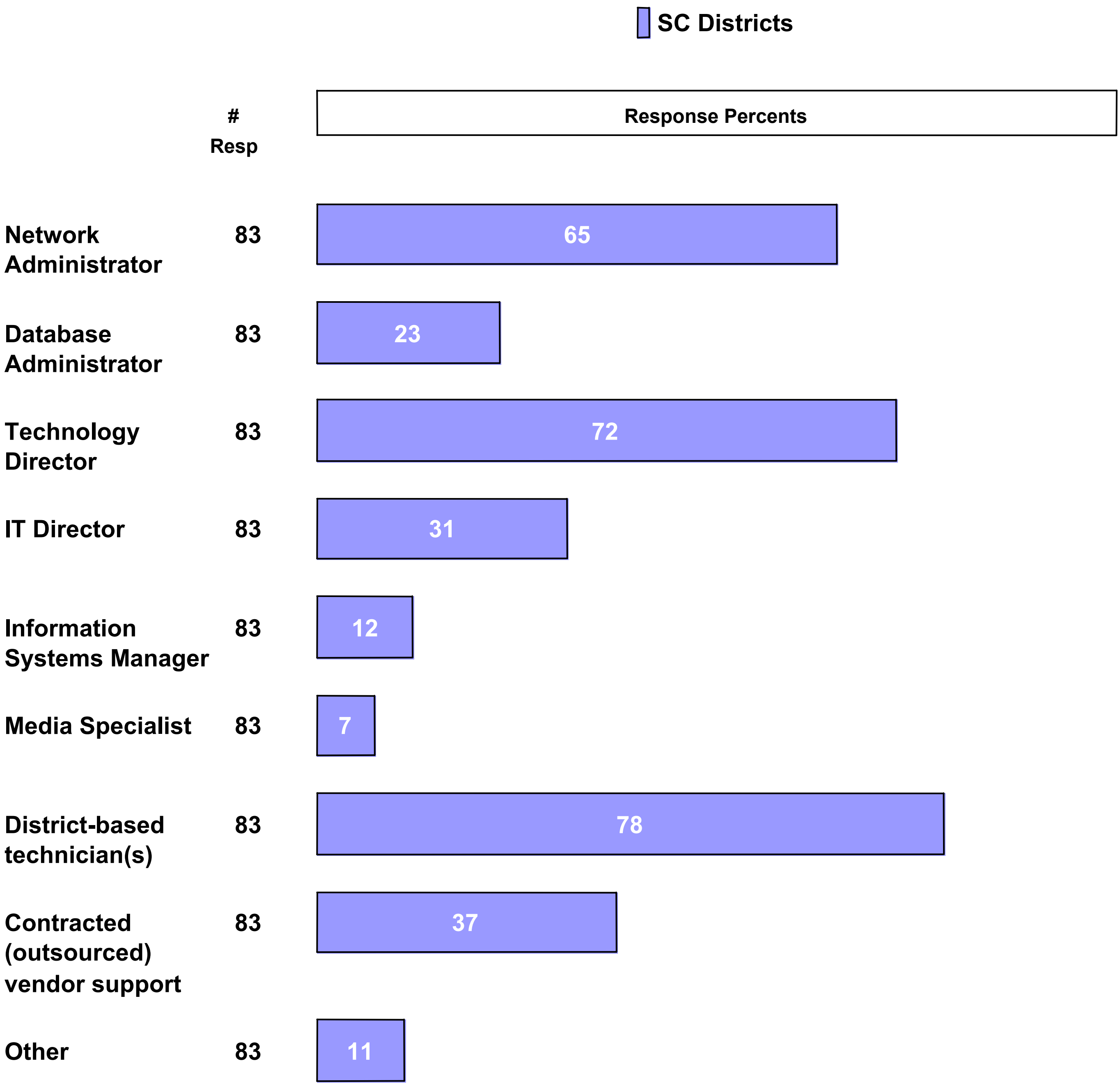
Please indicate the number of Information Technology (IT) staff available at your district level.



The mean for the State was 5.5.

District Staffing/Training

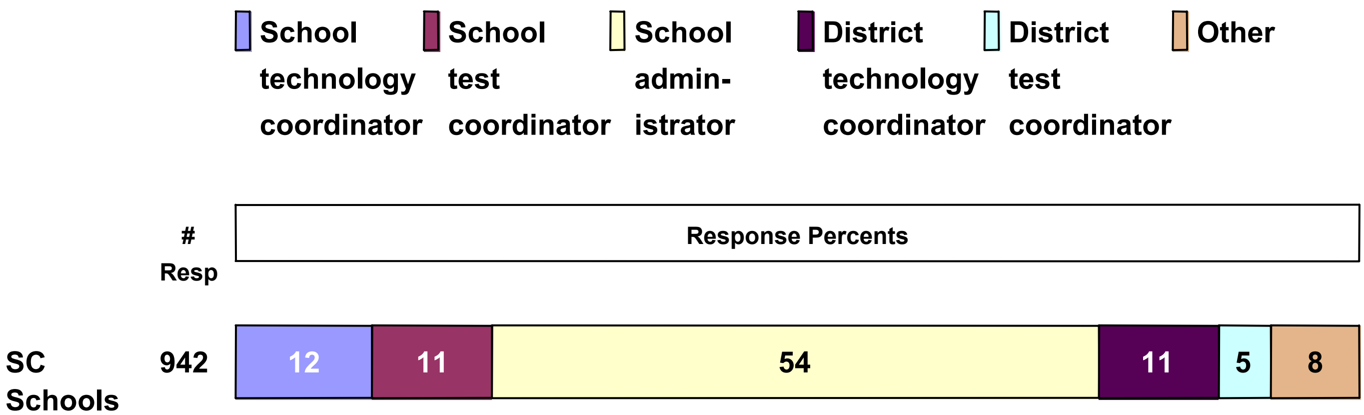
What types of IT personnel are available at your district level?



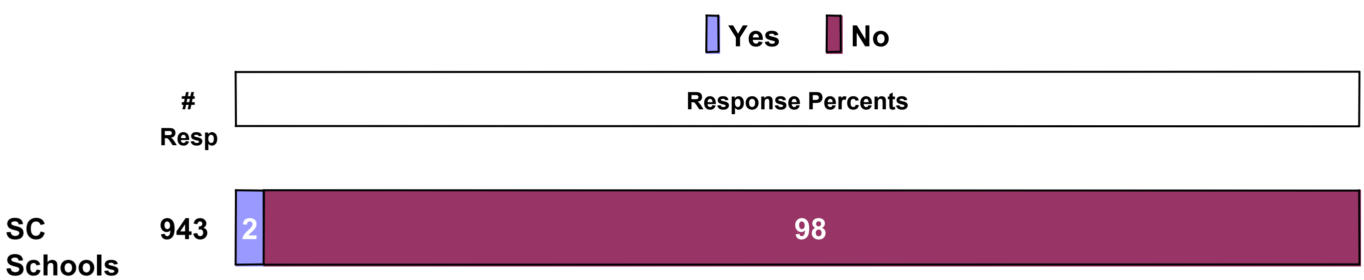
School Survey Results

School Demographics

Please indicate the title that best fits your role at this school.

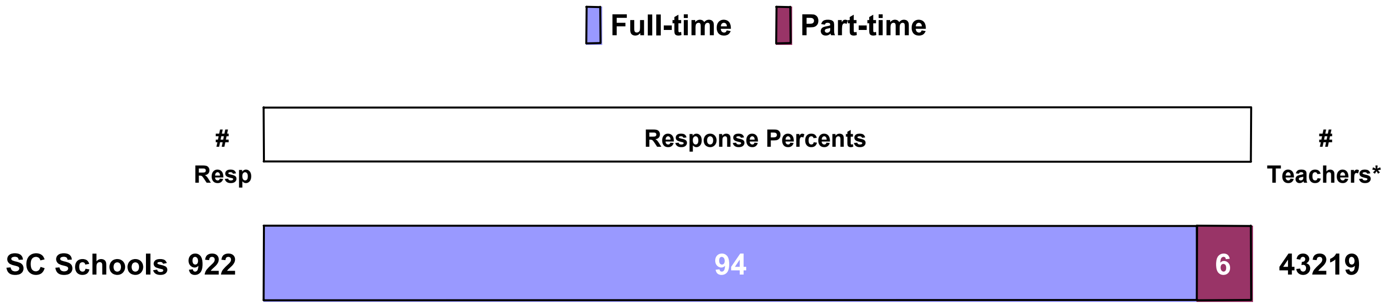


Is your school considered an adult education center?

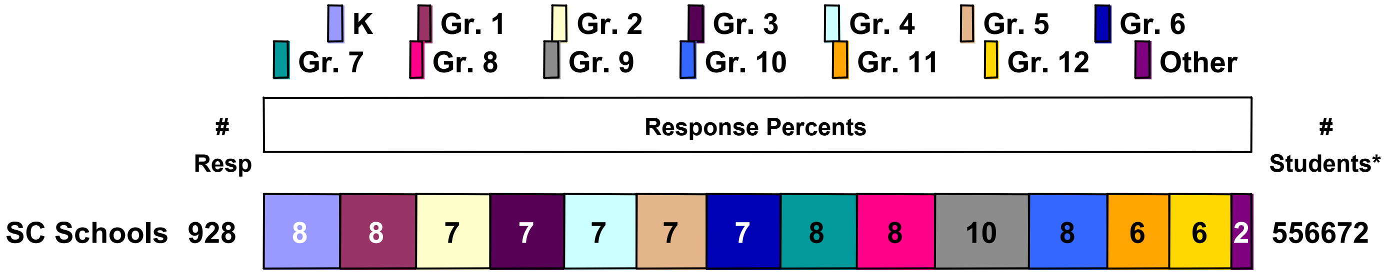


School Demographics

What is the total number of full-time/part-time teachers/instructional staff at your school?



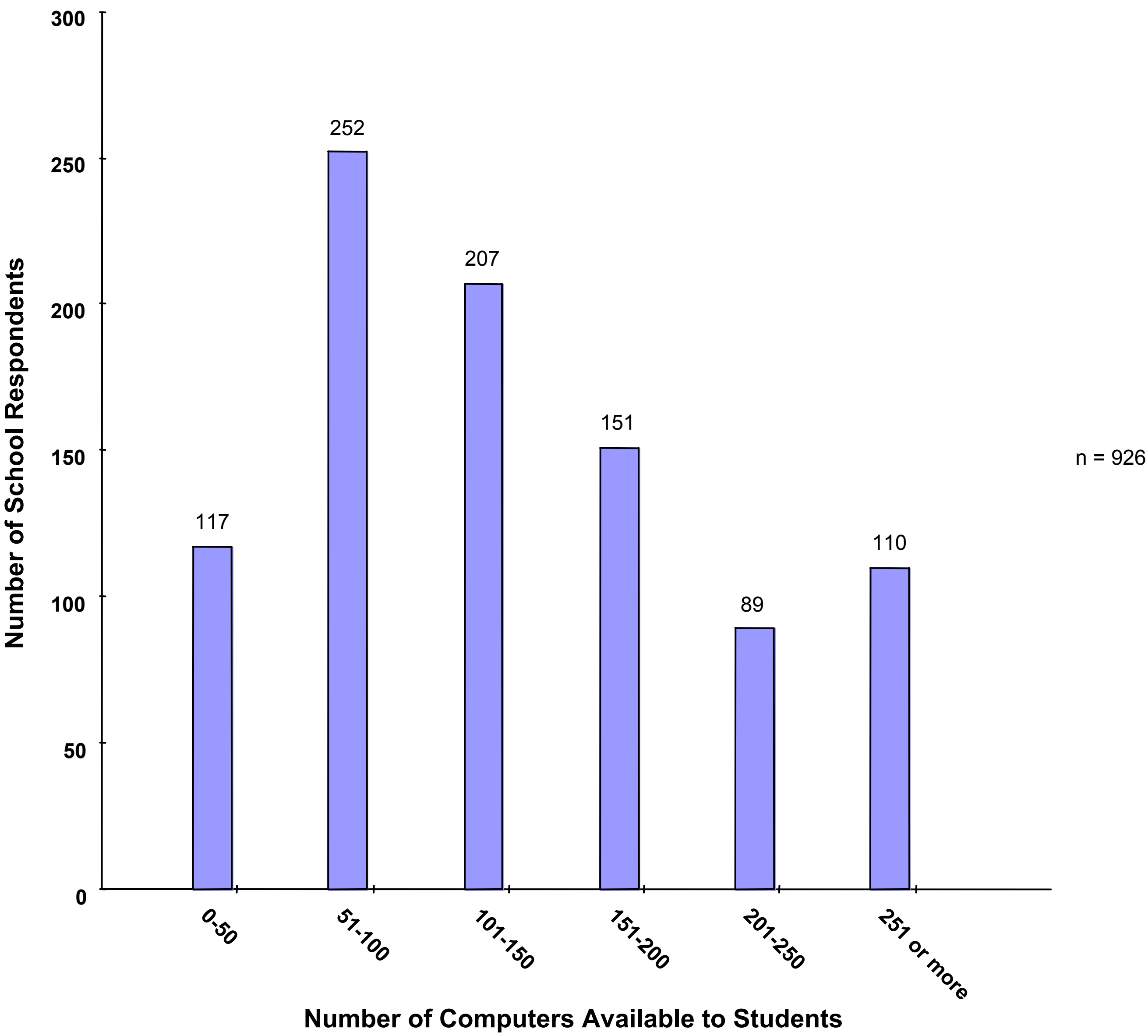
What is the total number of students in your school for each grade?



* The total number of staff/students reflects numbers reported by survey respondents.

School Hardware/Software

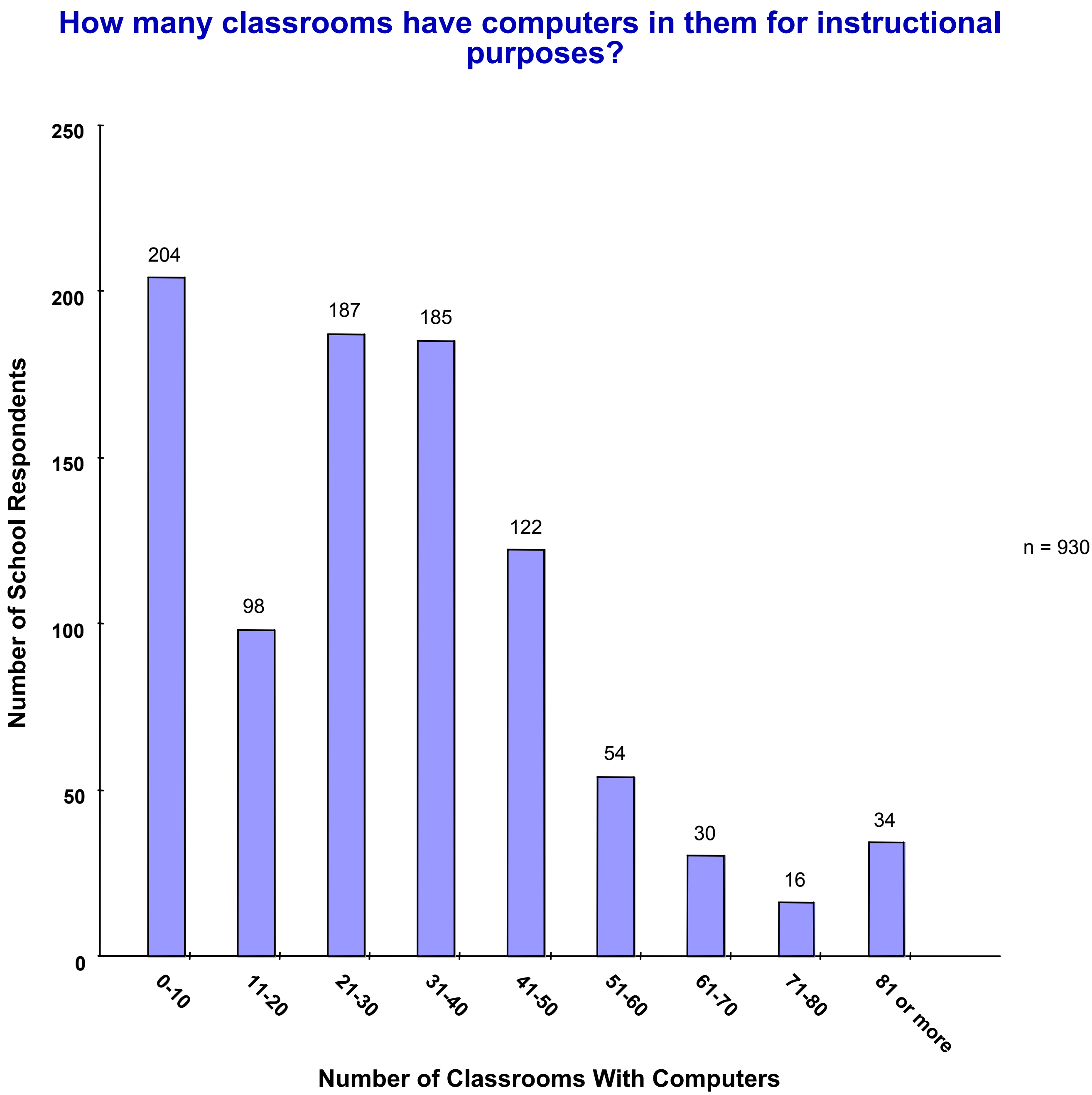
Please indicate the total number of computers available to students at your school.



Student-to-Computer Ratios

The student-to-computer ratio for SC Schools was 4:1

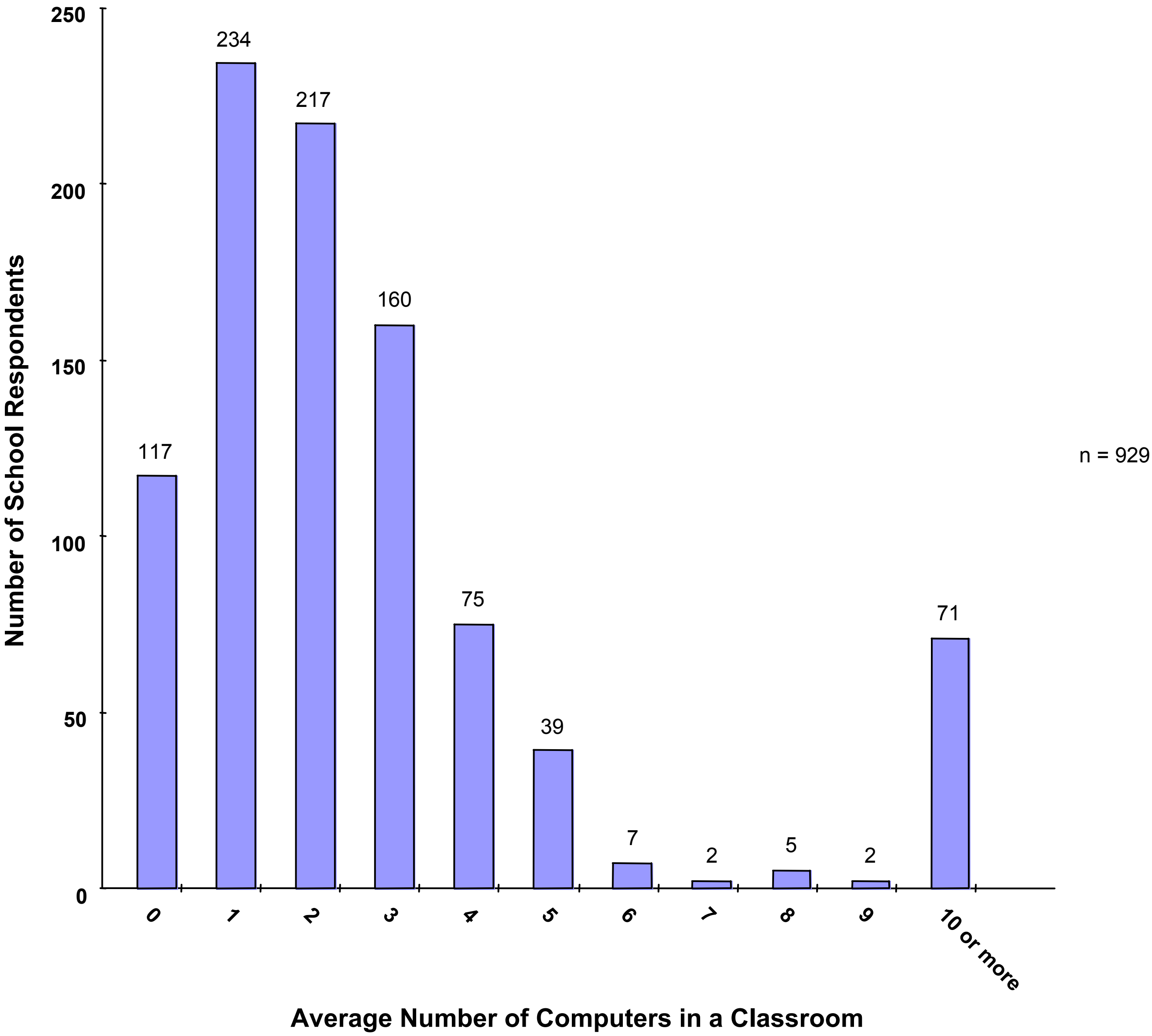
School Hardware/Software



The mean for the State was 31.7.

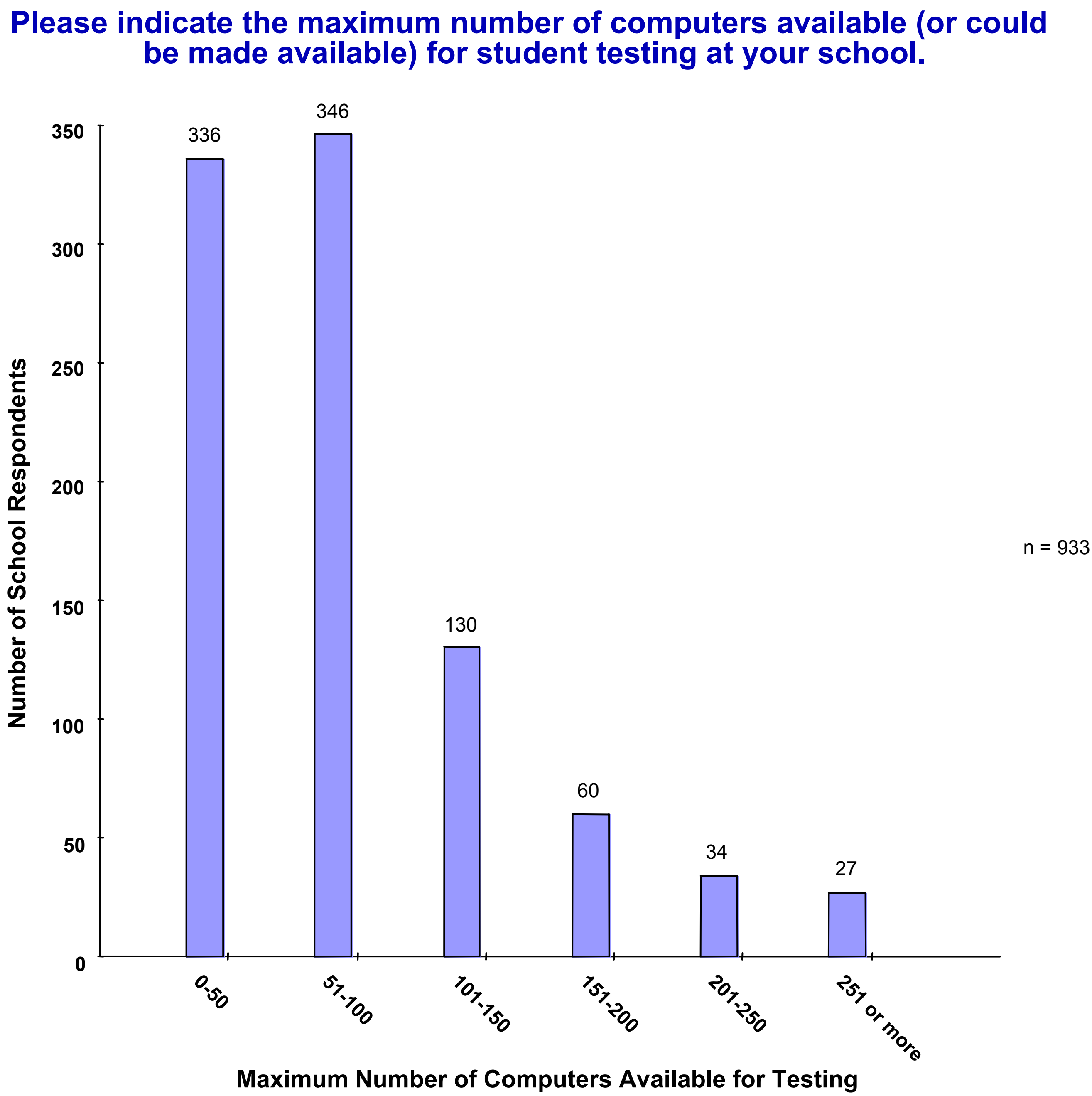
School Hardware/Software

On average, how many computers are in use for student instruction in each classroom?



The mean for the State was 6.0.

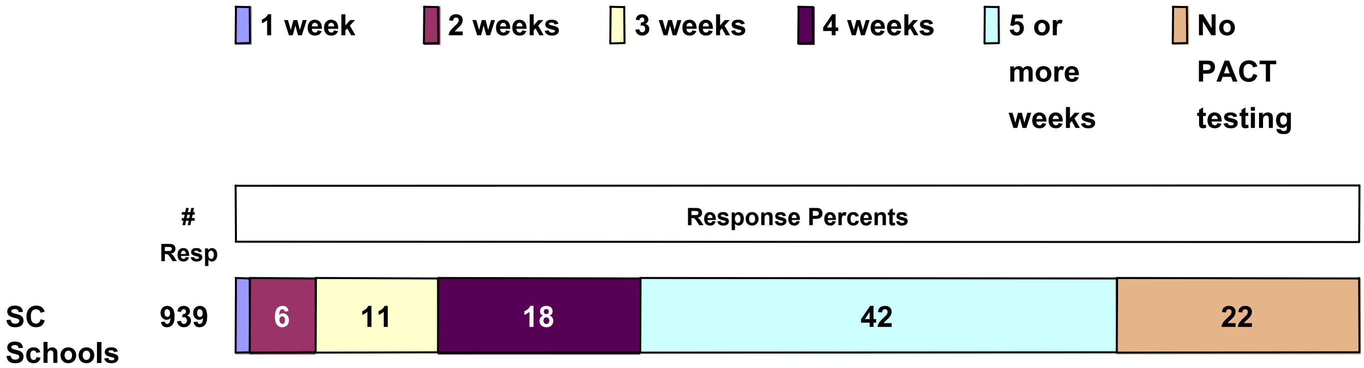
School Hardware/Software



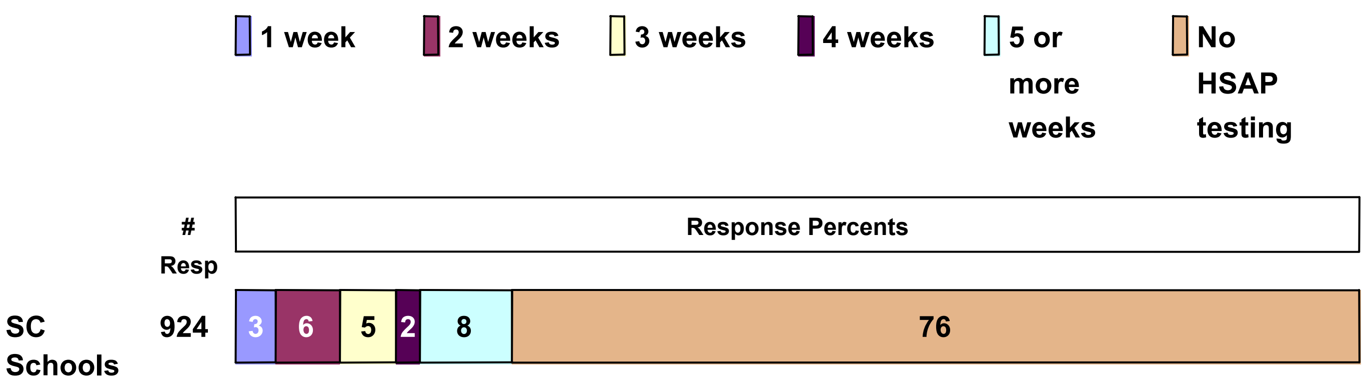
The mean for the State was 83.7.

School Space for Computerized Testing

My school has enough computers for all students to complete PACT (grades 3-8) testing if the testing window is:

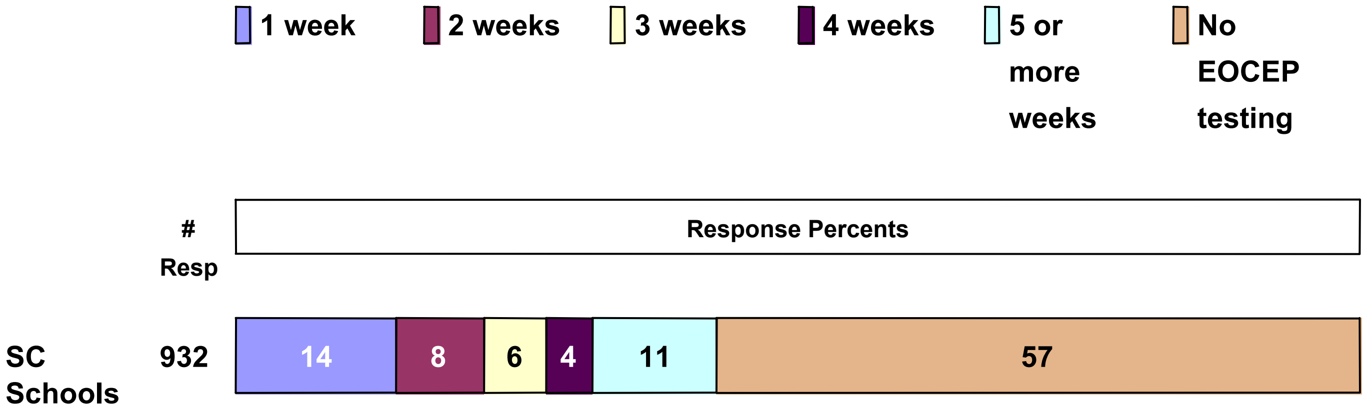


My school has enough computers for all students to complete HSAP (grades 10, 11, 12) testing if the testing window is:

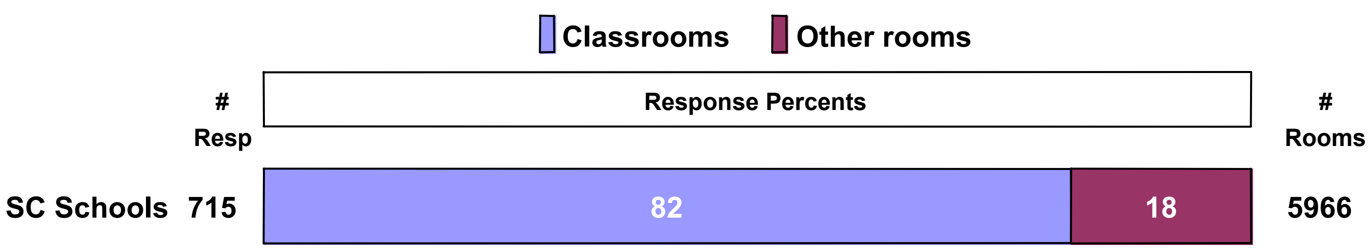


School Space for Computerized Testing

My school has enough computers for all students to complete EOCEP (grades 7-12) testing if the testing window is:

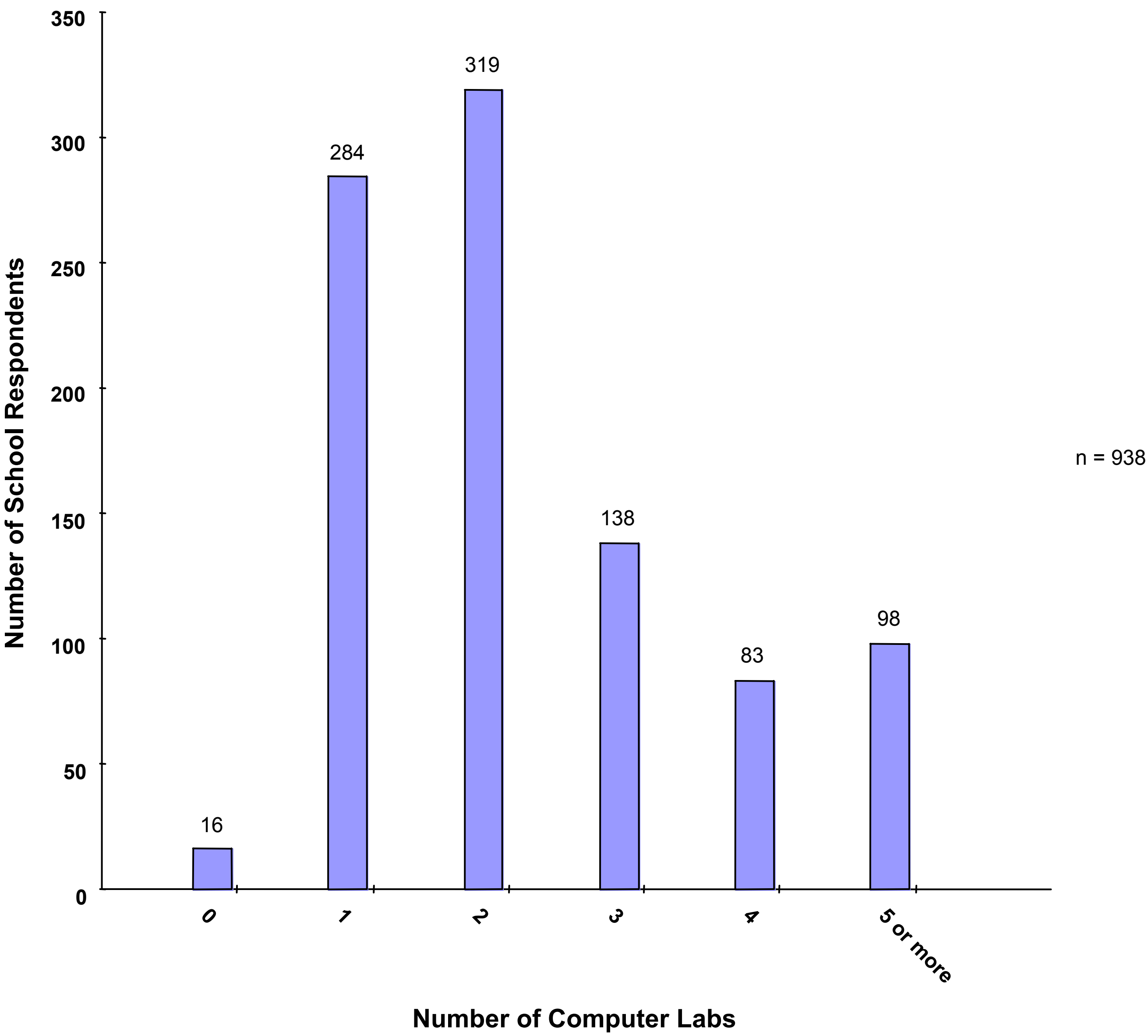


How many classrooms/other rooms could be used for student testing?



School Space for Computerized Testing

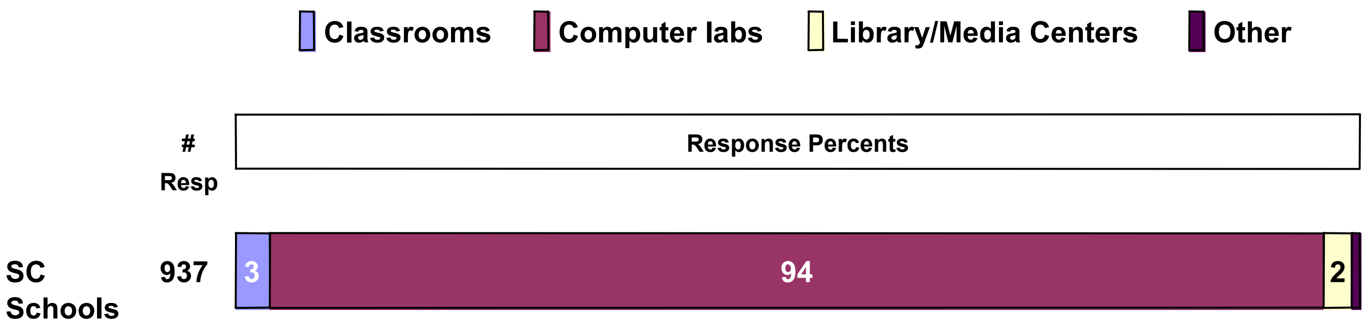
Please indicate the number of computer labs in your school, including mobile labs.



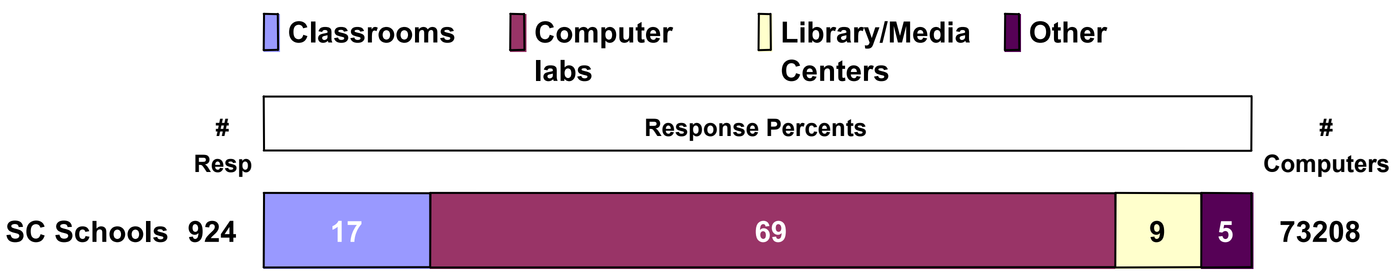
The mean for the State was 2.5.

School Computer Locations

In which area would students most likely be given computerized tests?

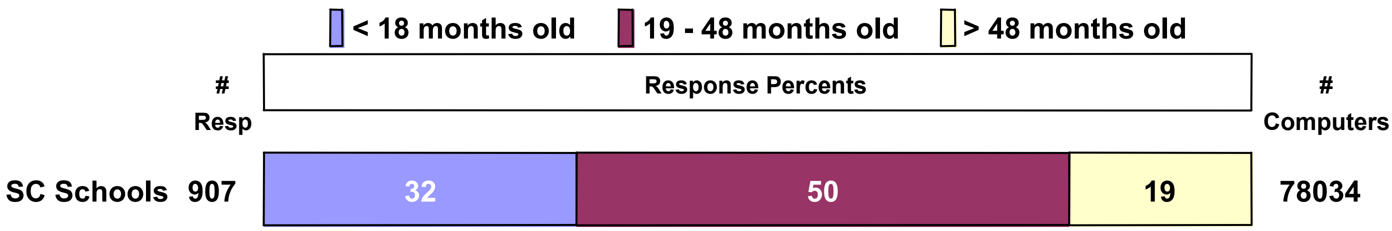


For computers that would be used for student testing, please list the number of computers located in each of the following areas:



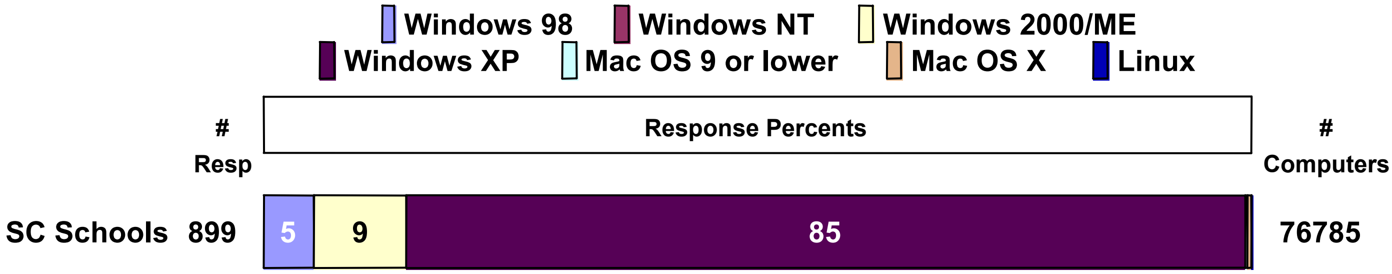
School Equipment Age Distribution

For computers that would be used for student testing, please list the number of computers:



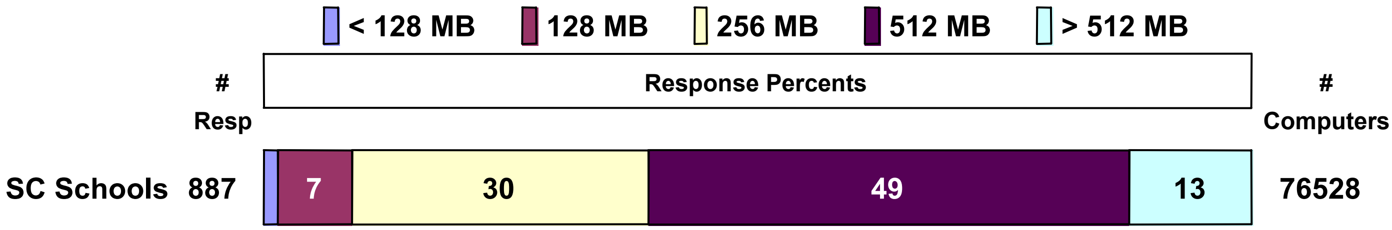
School Equipment OS Type Distribution

For computers that would be used for student testing, please list the number of computers with the following operating systems:



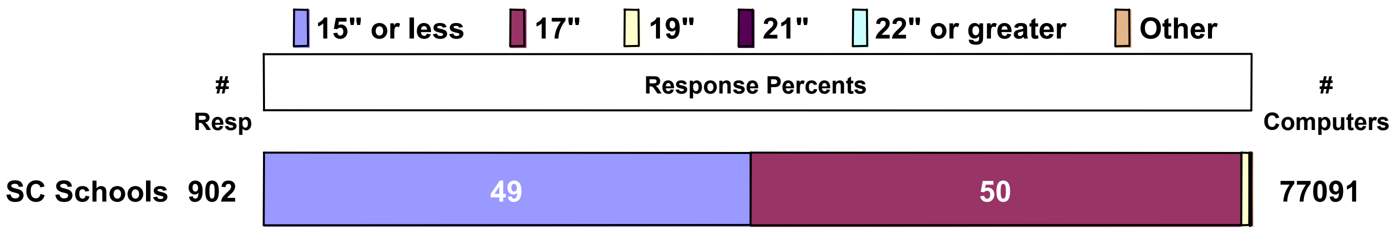
School Caching Capacity

For computers that would be used for student testing, please list the number of computers with the following amounts of internal memory (RAM):

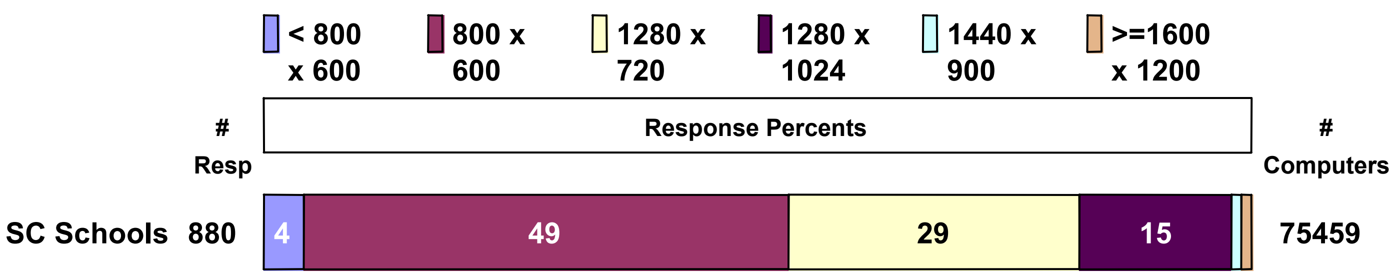


School Monitor Displays

For computers that would be used for student testing, please list the number of computers with a monitor size of:

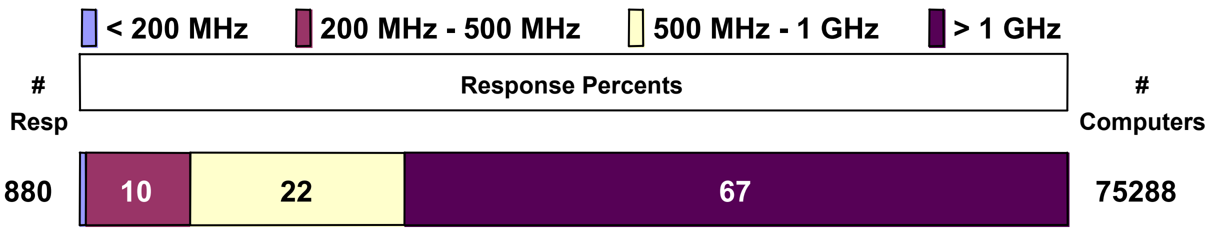


For computers that would be used for student testing, please list the number of computers with a monitor resolution of:

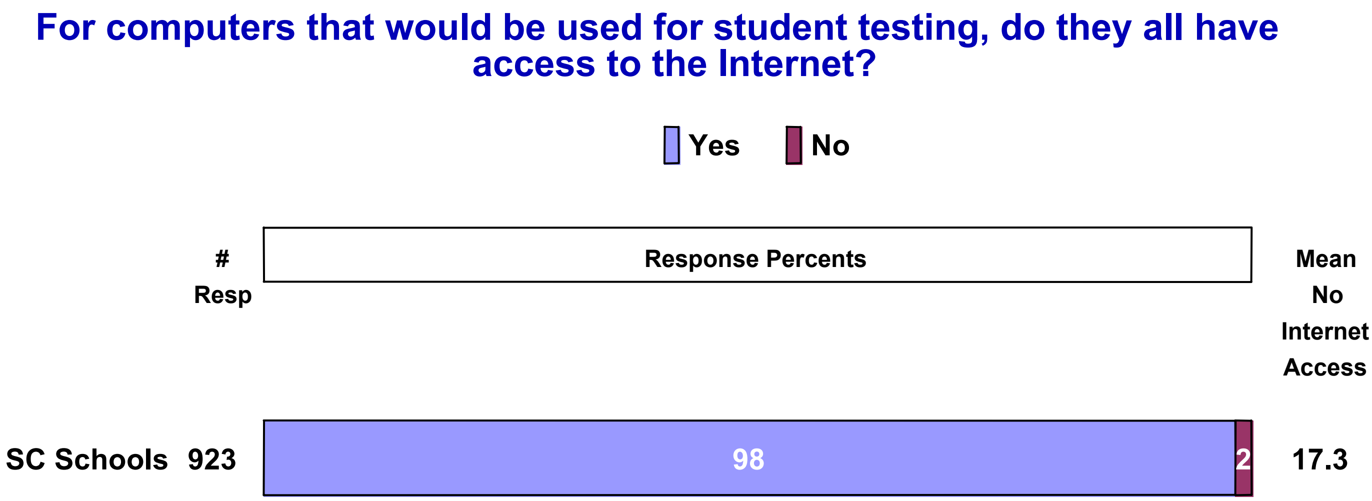


School CPU Processors

For computers that would be used for student testing, please list the number of computers with the following processor speeds:

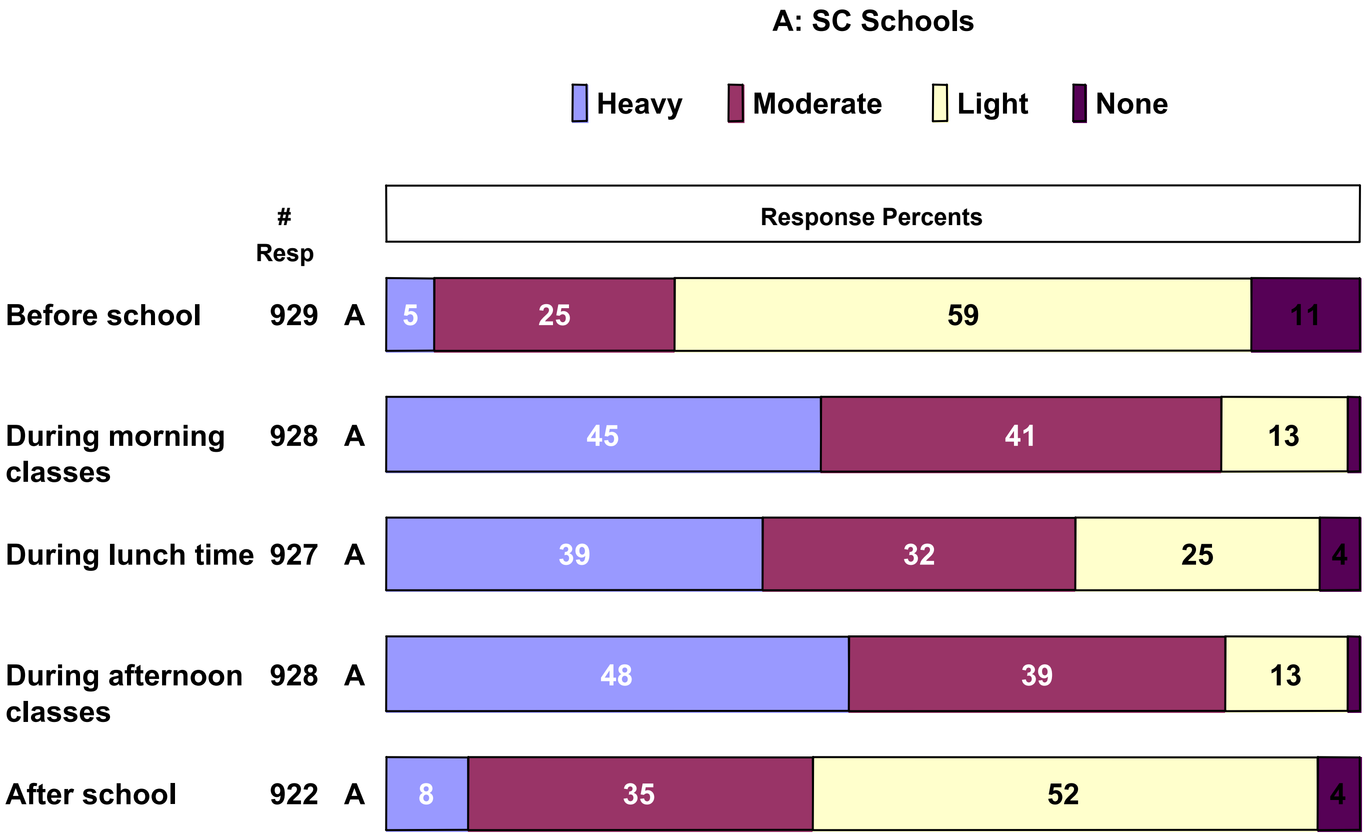


School Bandwidth



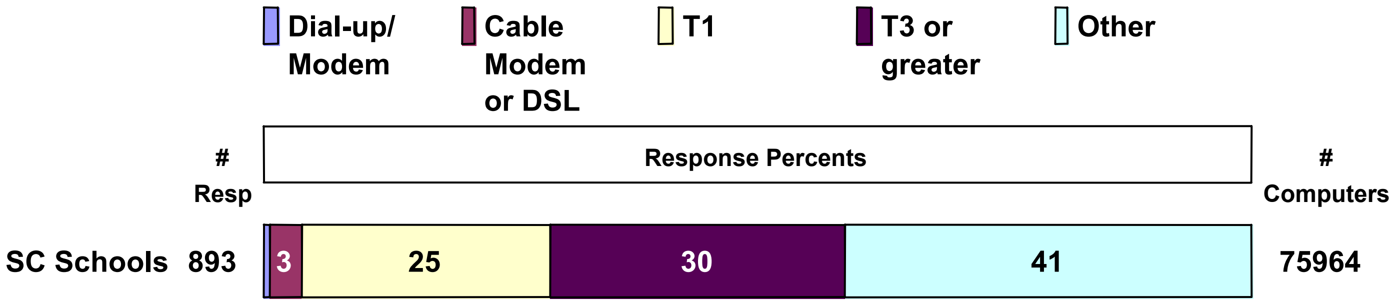
School Bandwidth

Approximately how would you describe overall Internet usage at your school at different times of the day?

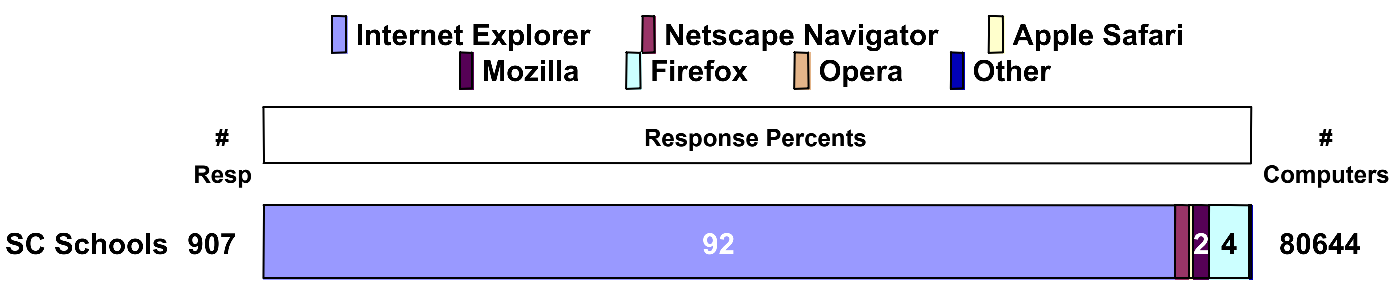


School Bandwidth

For computers that would be used for student testing, please estimate the number of computers connected to the Internet through the following means:

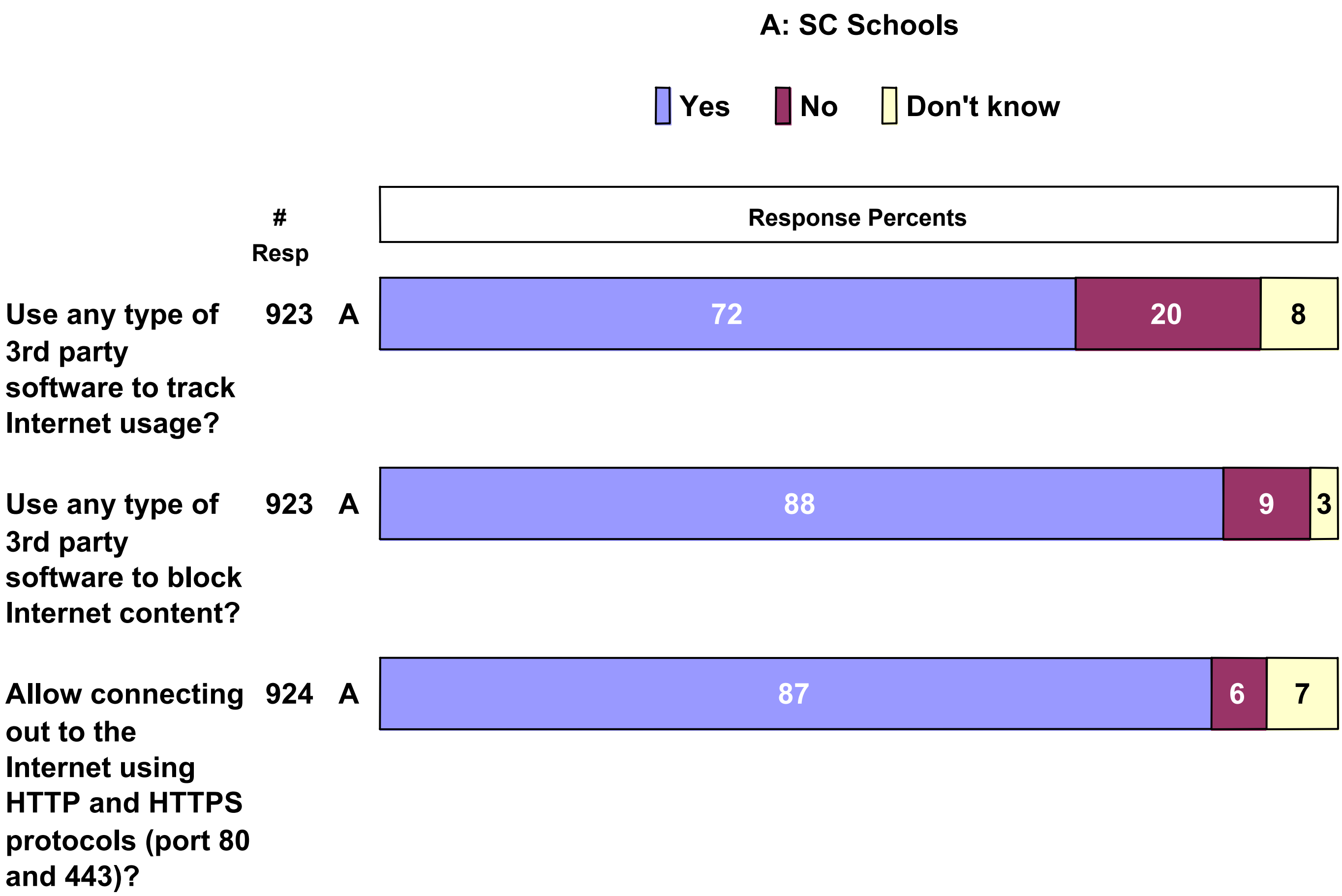


Of the computers that would be used for student testing, please estimate the number that use the following browsers for Internet access:



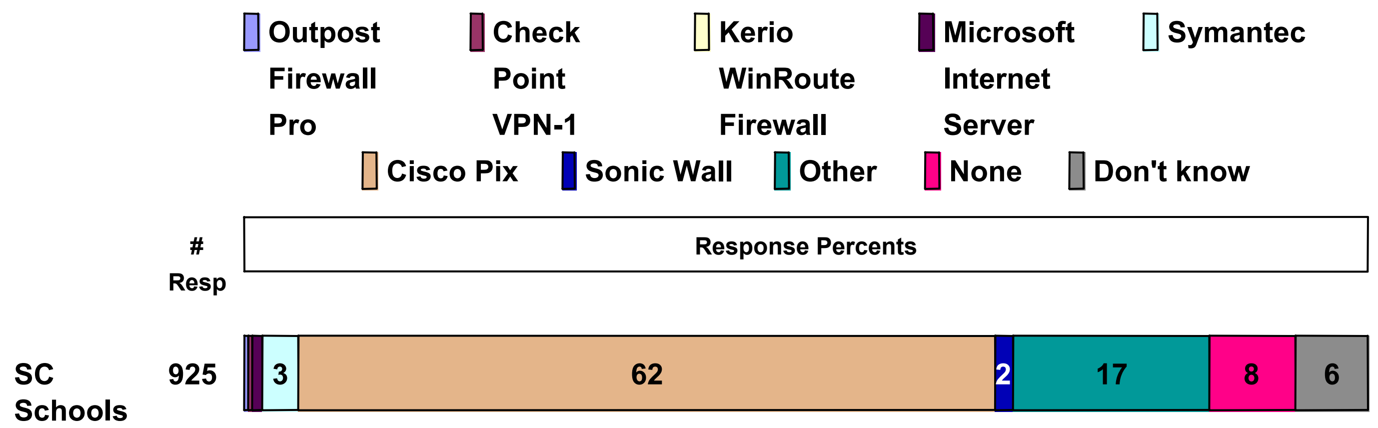
School Internet Monitoring/Filtering

For computers that would be used for student testing, does your school:



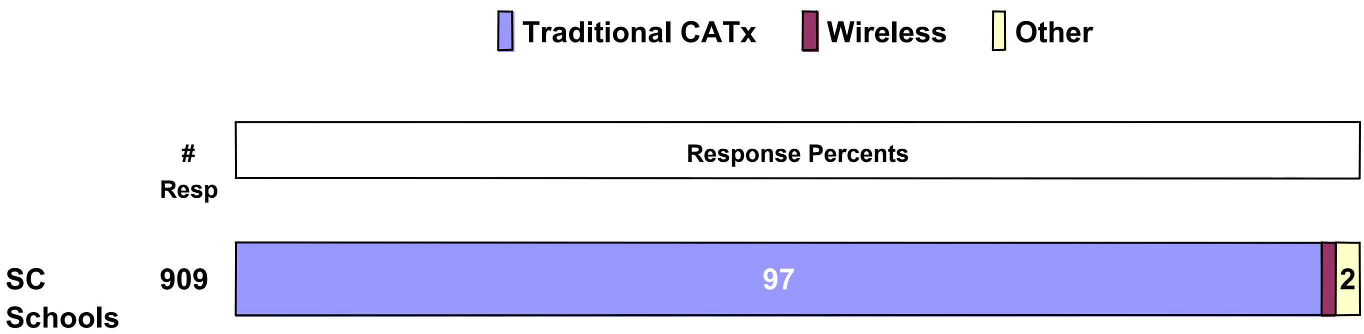
School Firewalls

**For computers that would be used for student testing, does your school use any type of firewall product?
Indicate if the firewall product is one of the following.**

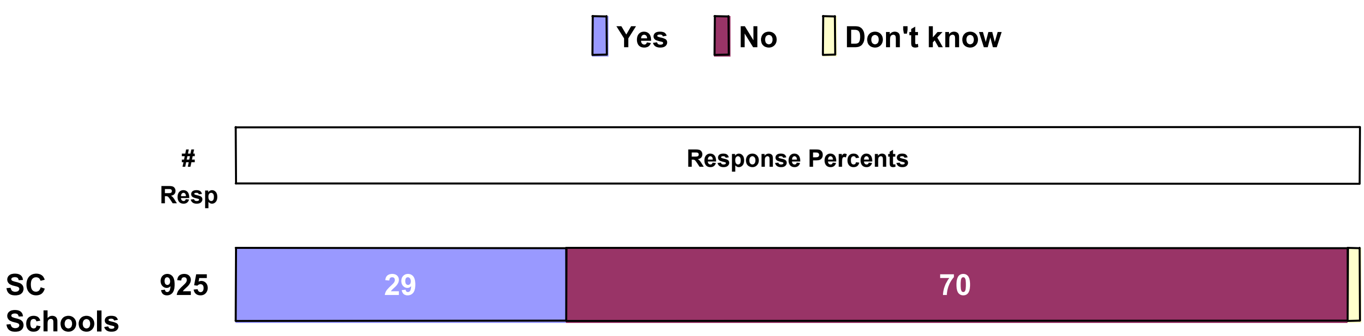


School Local Network Information

Select the connectivity type that best describes your location.

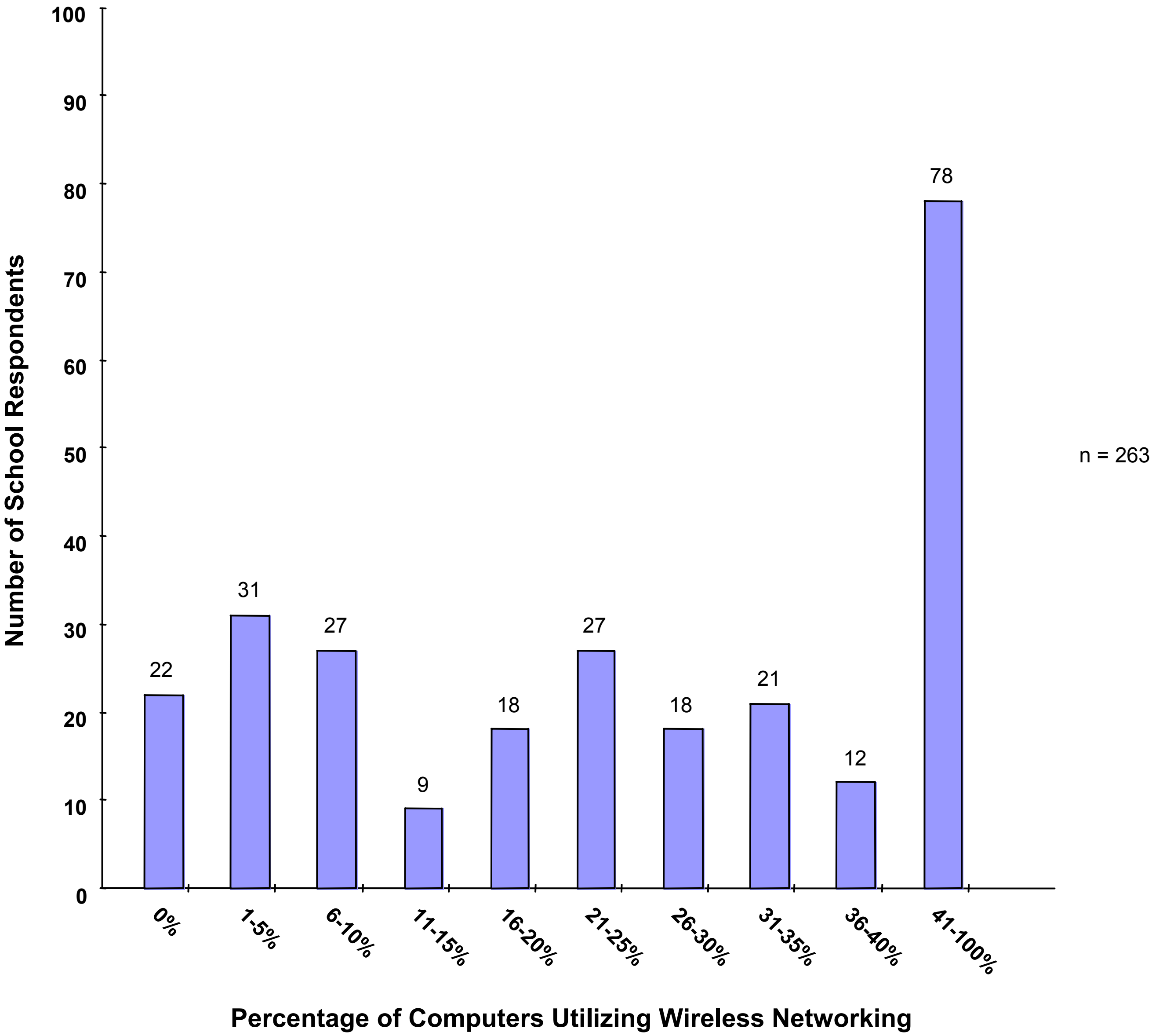


Do any of the computers that would be used for student testing use wireless networking for Internet access?



School Local Network Information

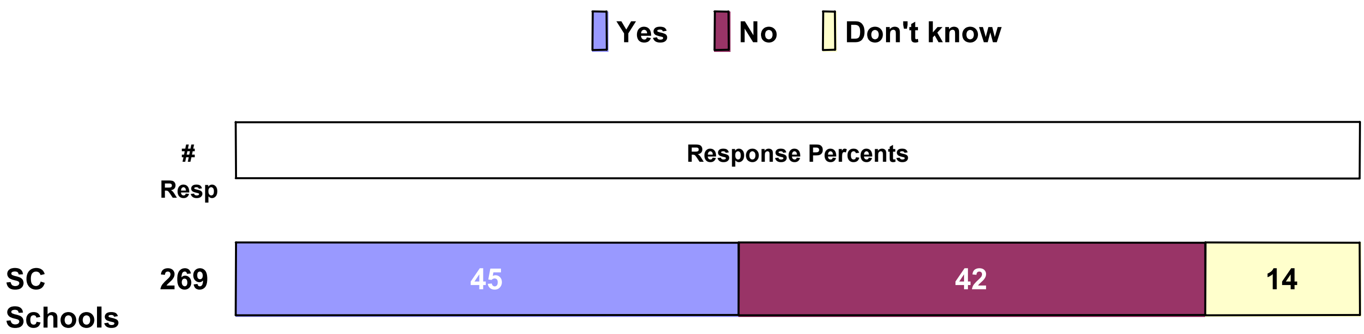
For computers that would be used for student testing, please indicate the approximate percentage of total computers in your school that utilize wireless networking.



The mean for the State was 28.5%.

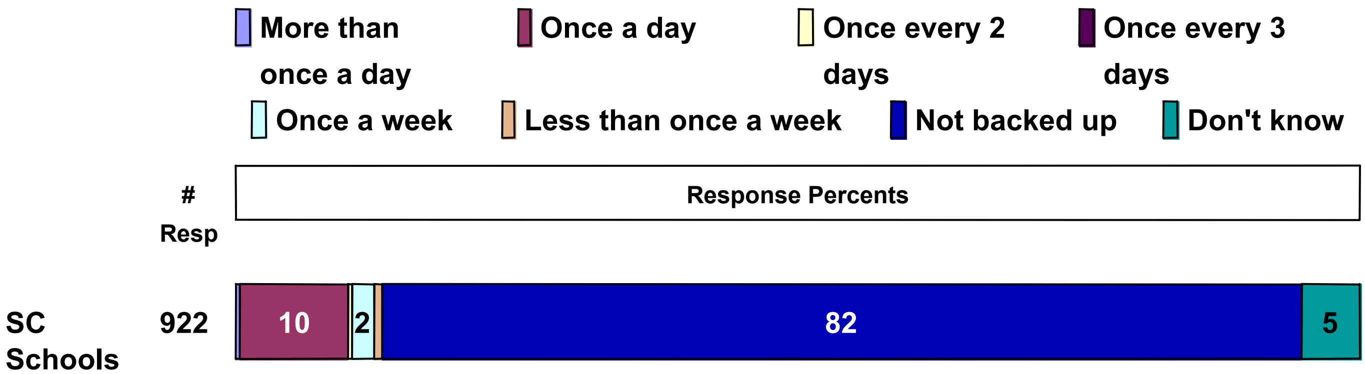
School Local Network Information

Is your wireless network using any form of encryption?

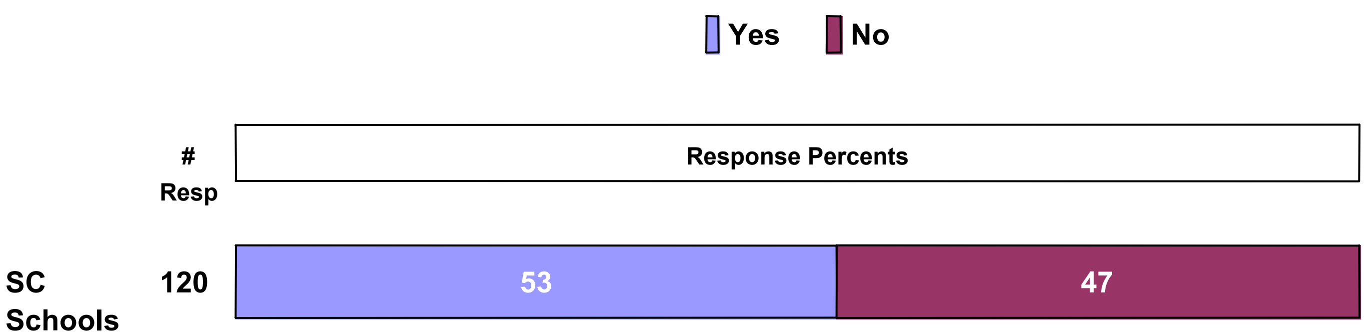


School Backup/Storage

For computers that would be used for student testing, are the hard drives of these computers backed up on a regular basis? How frequently are they backed up?

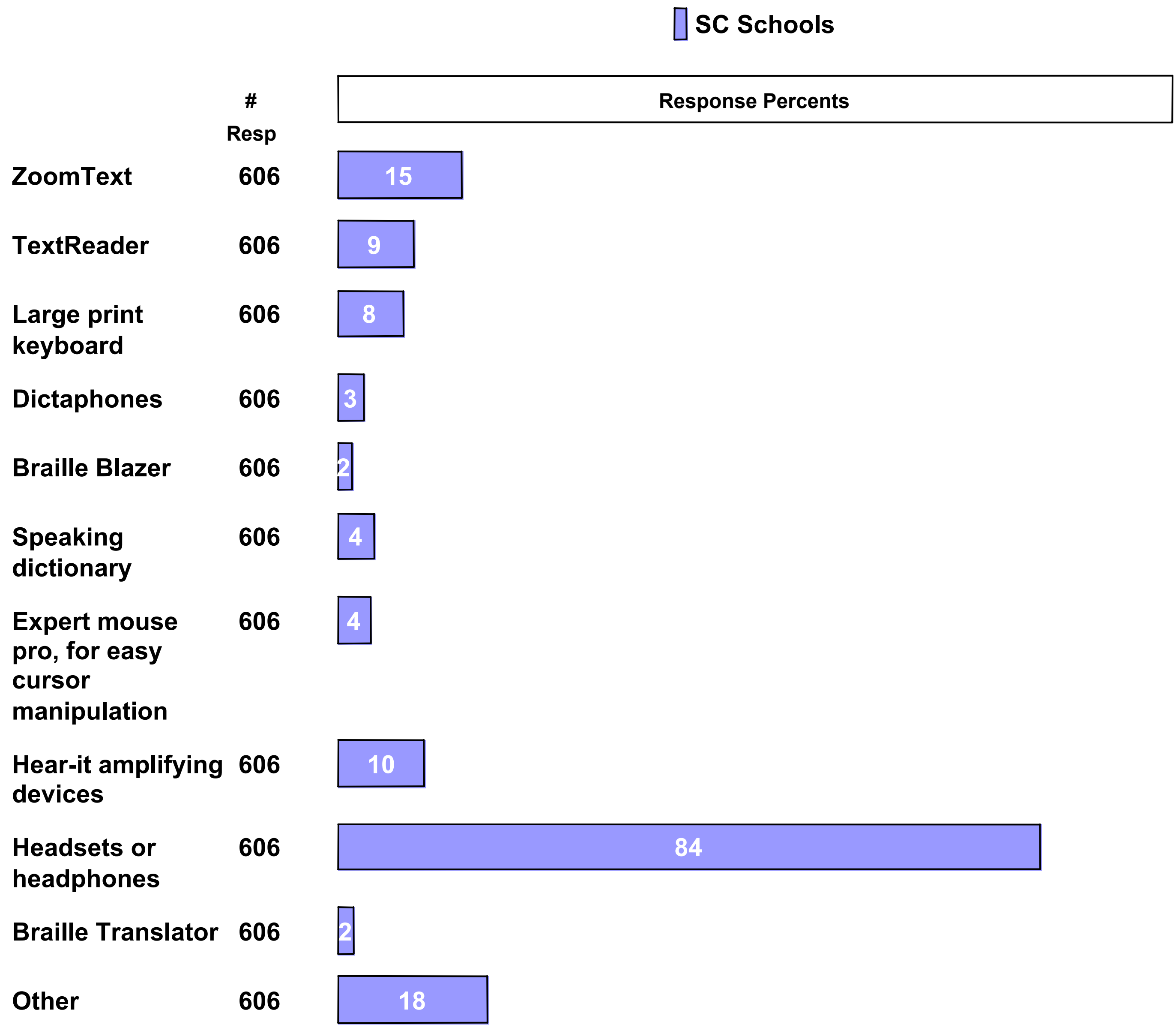


Are the backup media stored offsite?



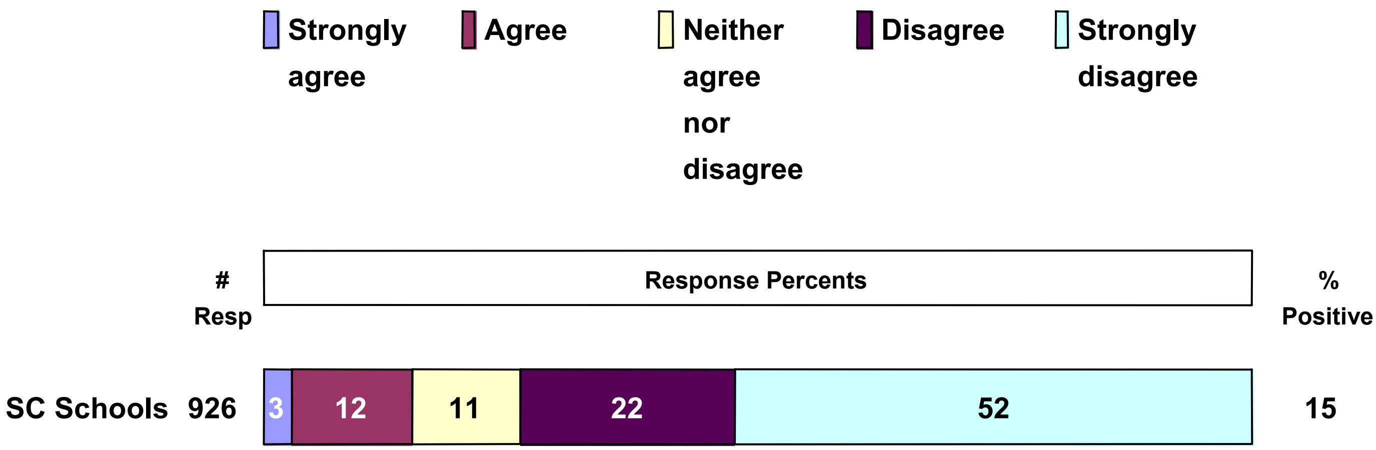
School Accessibility for Special Needs

Please indicate any specialized software/hardware available in your school to assist special needs students.

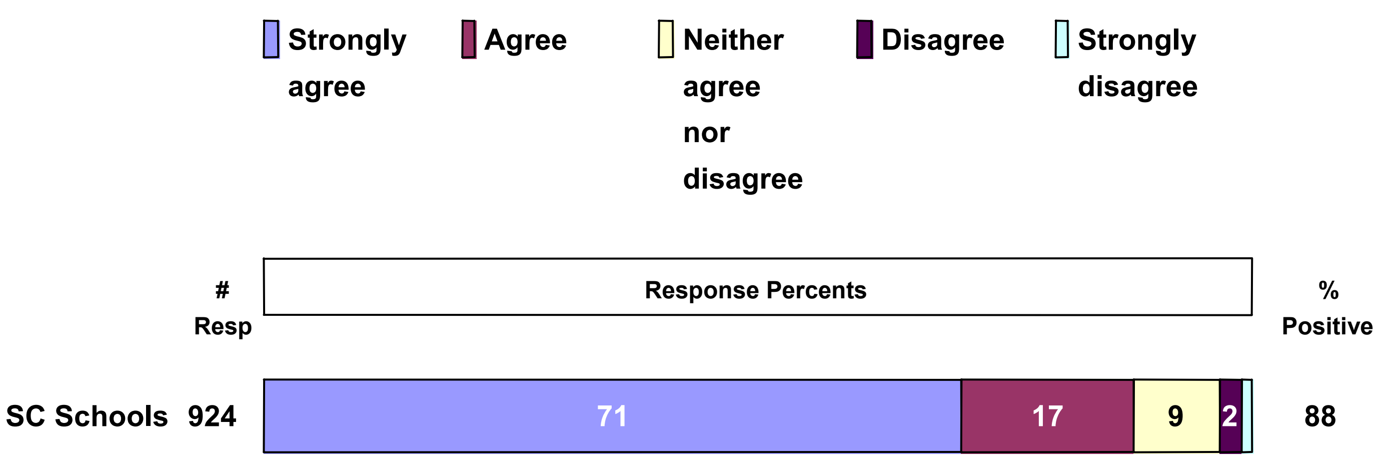


School Refresh Budgets

My school has an adequate budget to maintain computers that would be used for student testing.

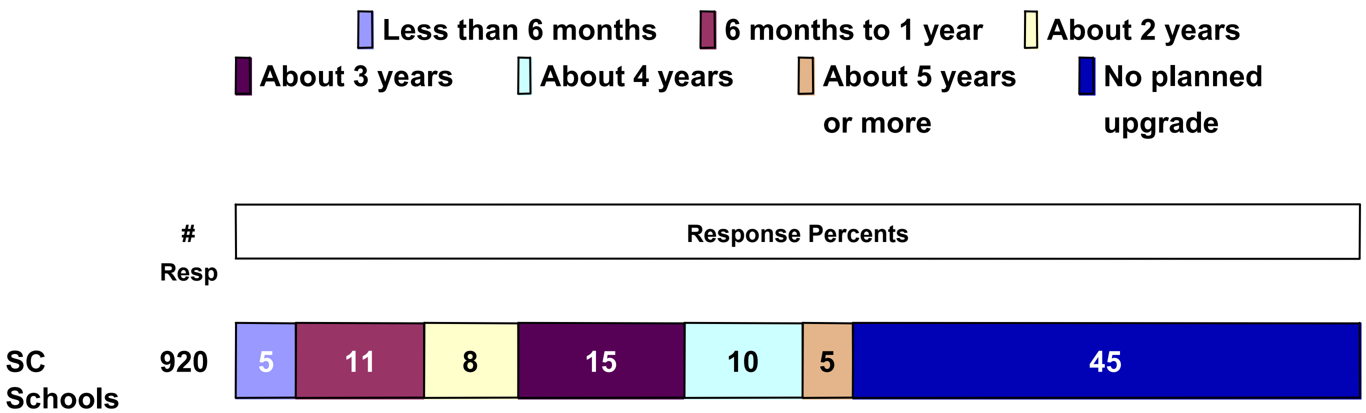


My school would require additional funding to acquire the computers necessary for student testing.



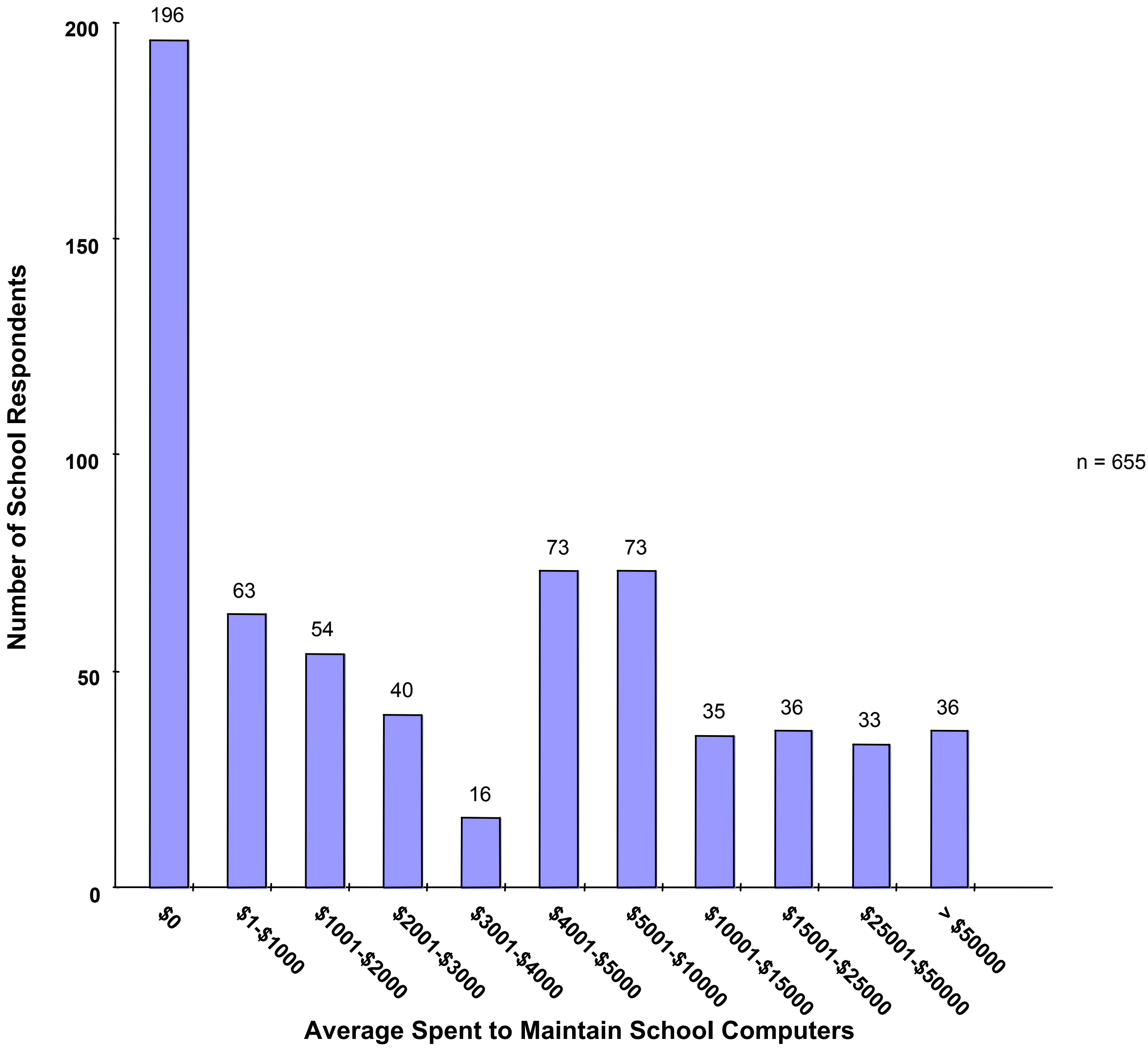
School Refresh Budgets

Do you have plans to upgrade your student computer fleet? When is the upgrade planned to take place?



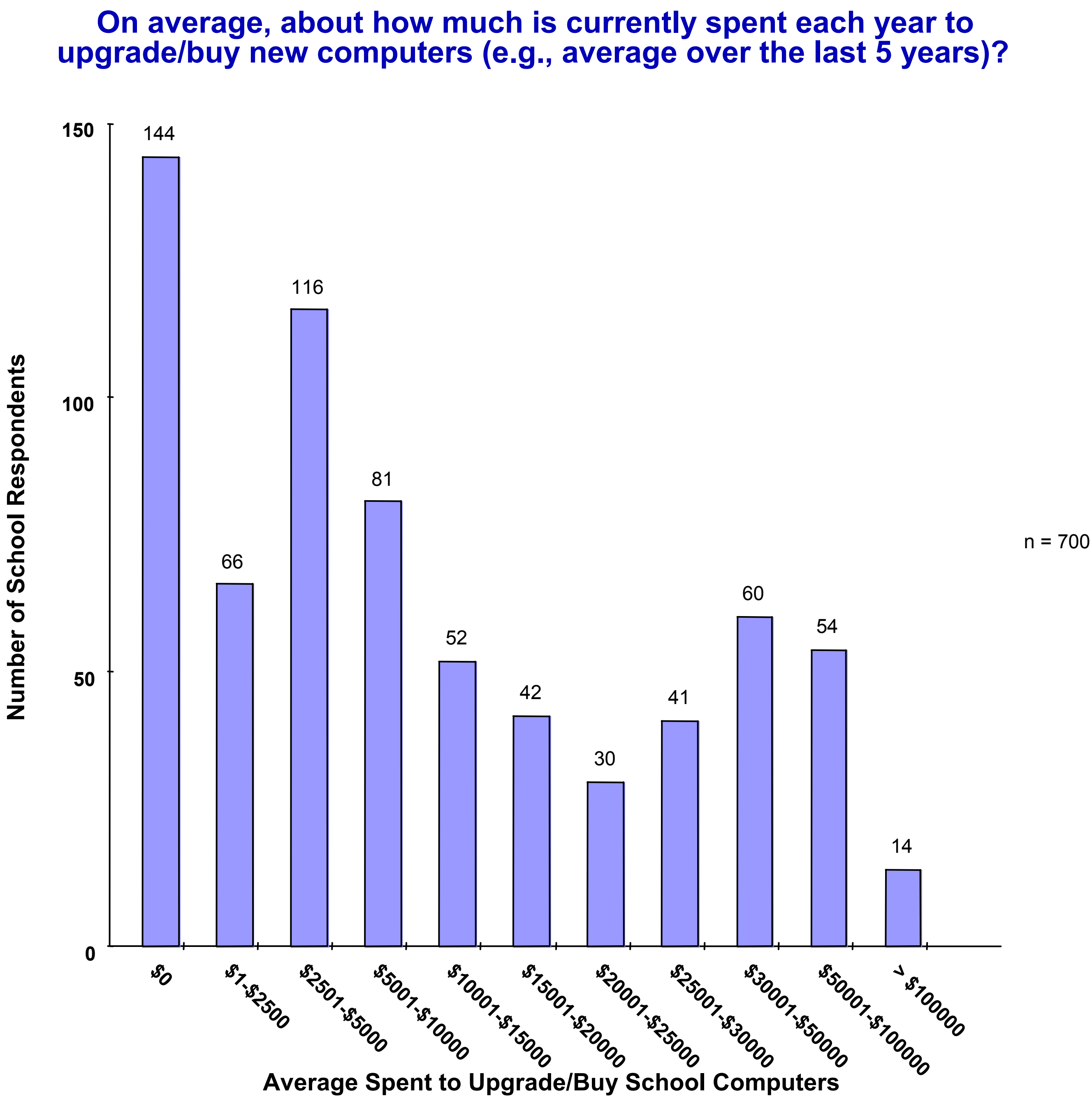
School Refresh Budgets

On average, about how much is currently spent each year to maintain your school computers (e.g., average over the last 5 years)?



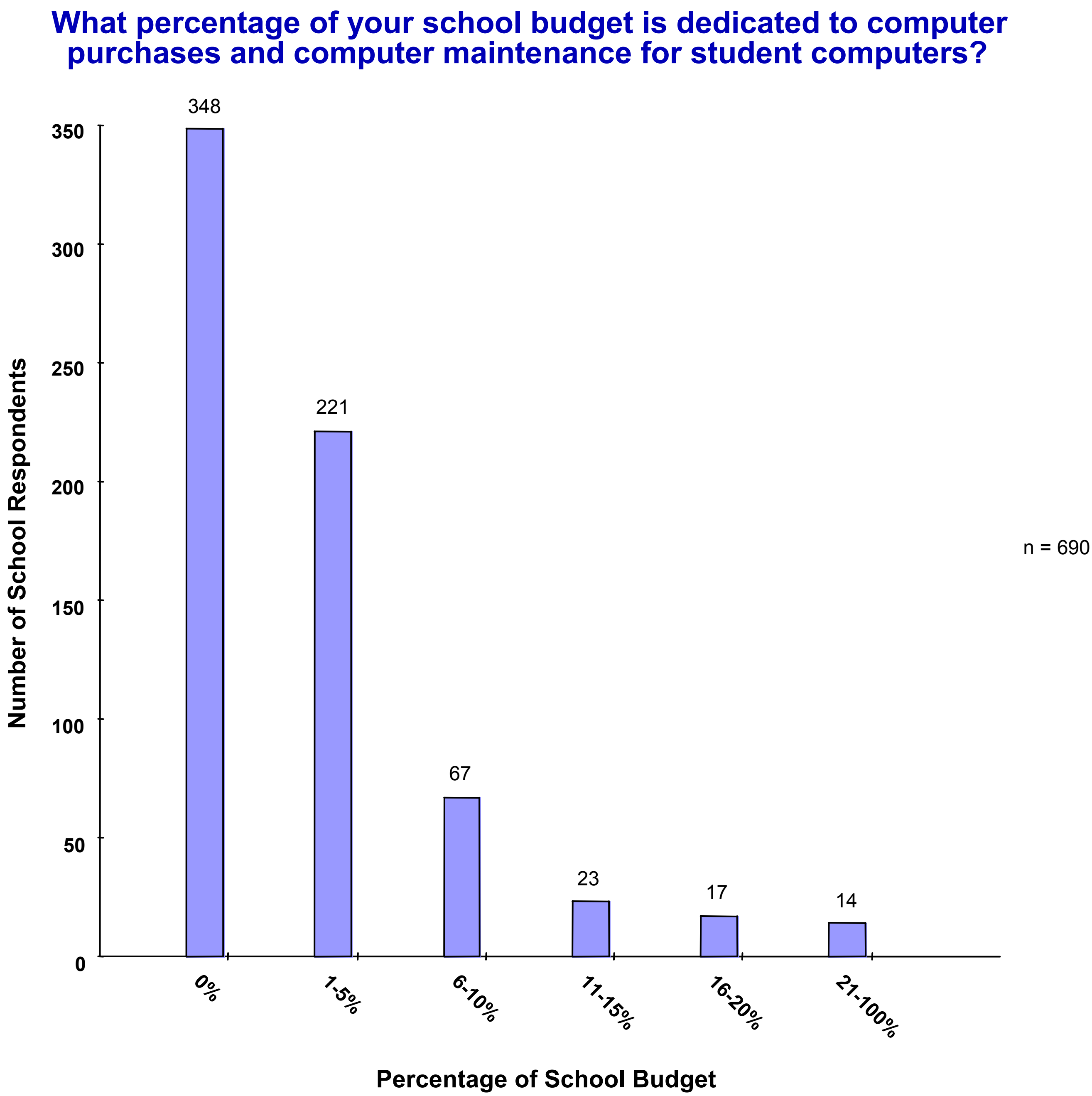
The mean for the State was $14951.69.

School Refresh Budgets



The mean for the State was $20480.45.

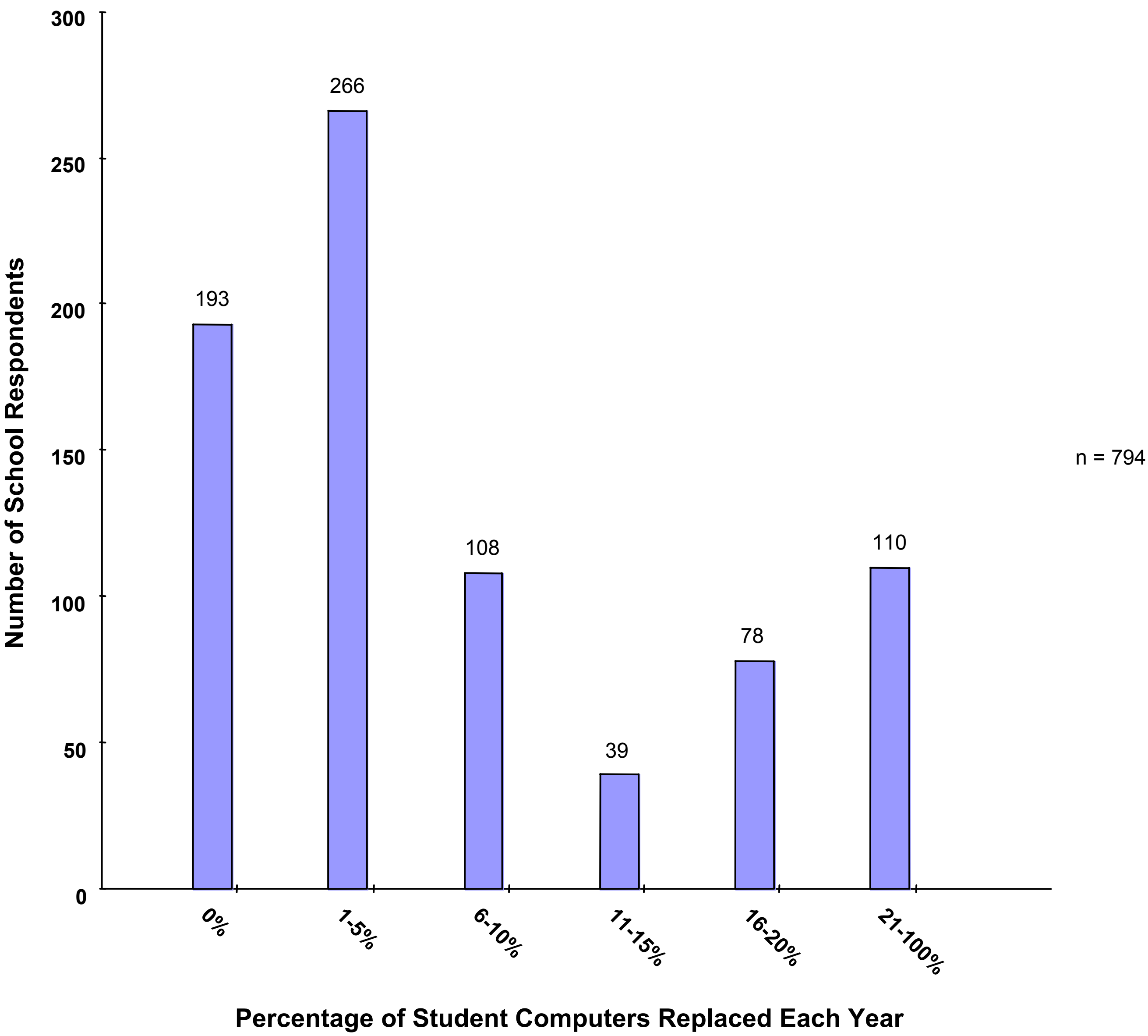
School Refresh Budgets



The mean for the State was 3.5%.

School Refresh Budgets

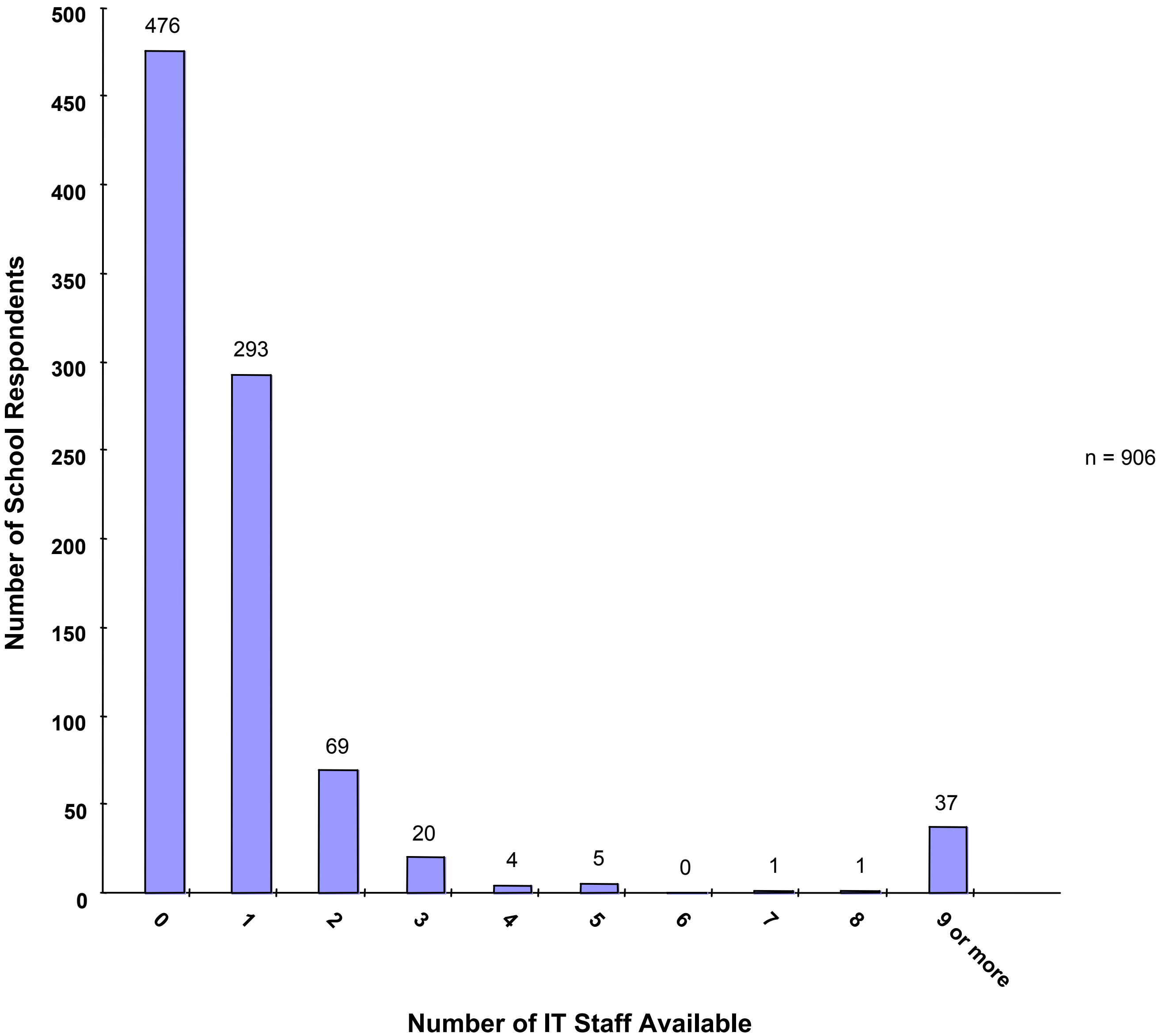
On average, what percentage of student computers in your school is replaced each year?



The mean for the State was 9.2%.

School Staffing/Training

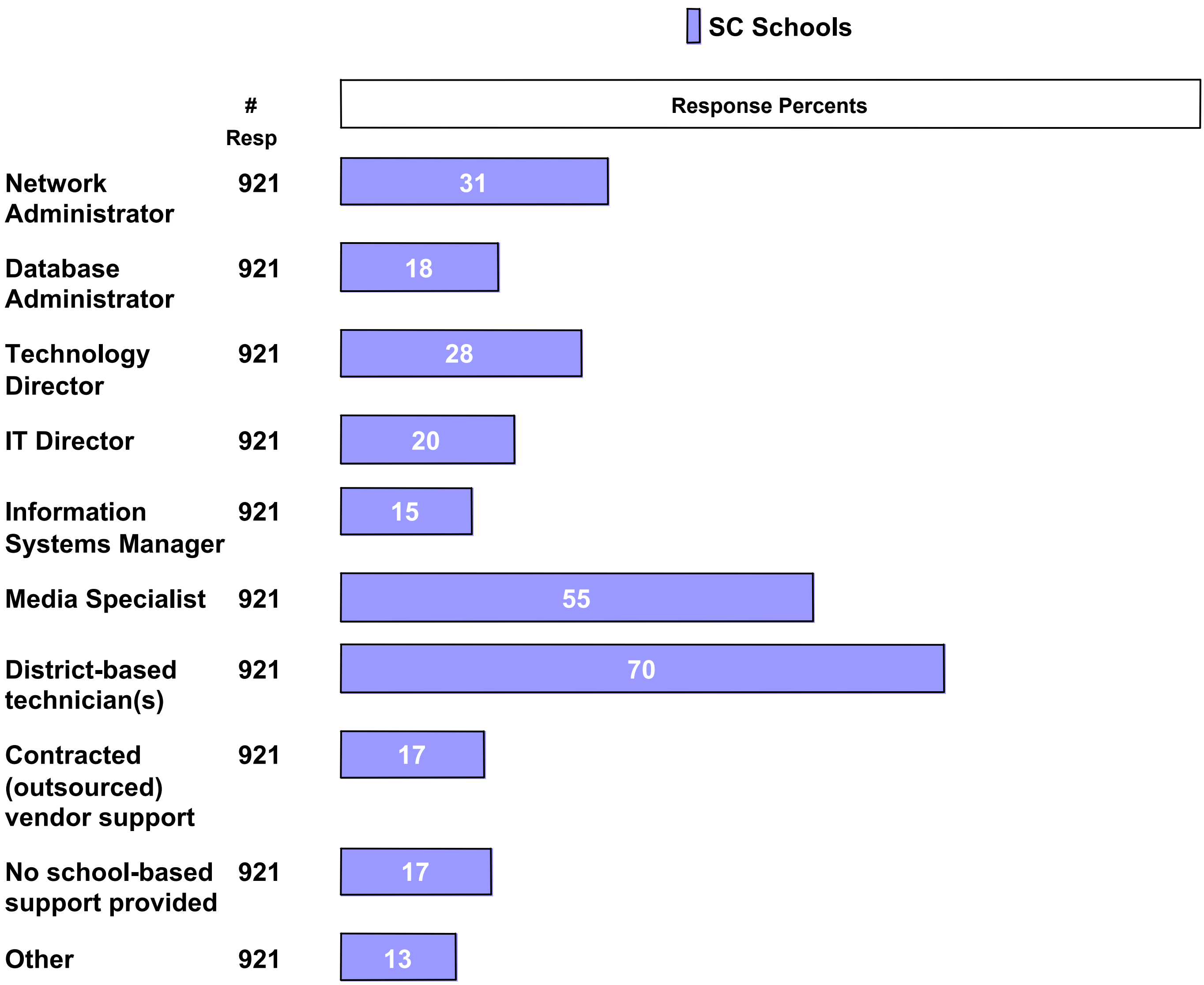
Please indicate the total number of Information Technology (IT) staff available at your school.



The mean for the State was 1.4.

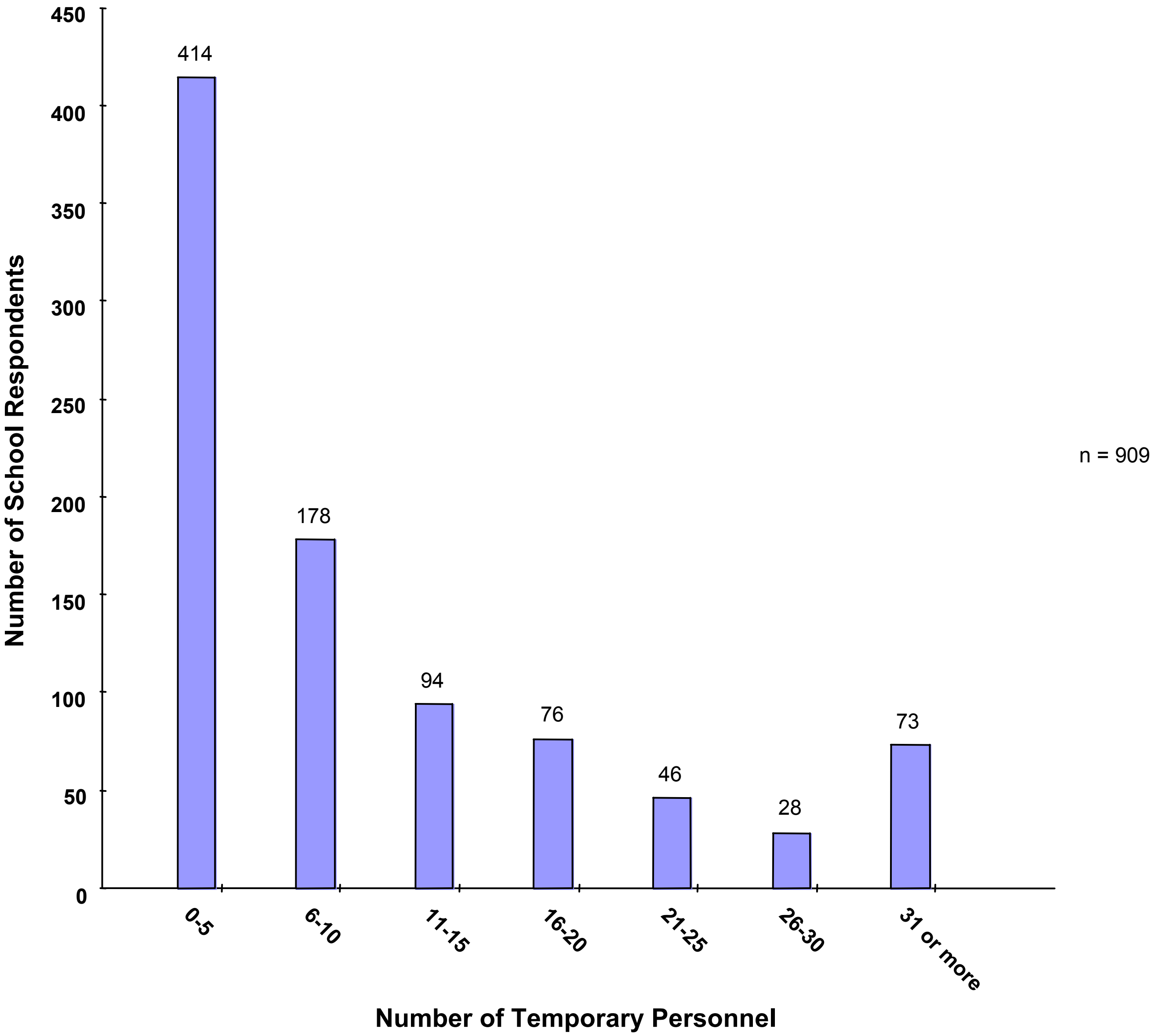
School Staffing/Training

What types of IT personnel are available for your school?



School Temporary Personnel Needed

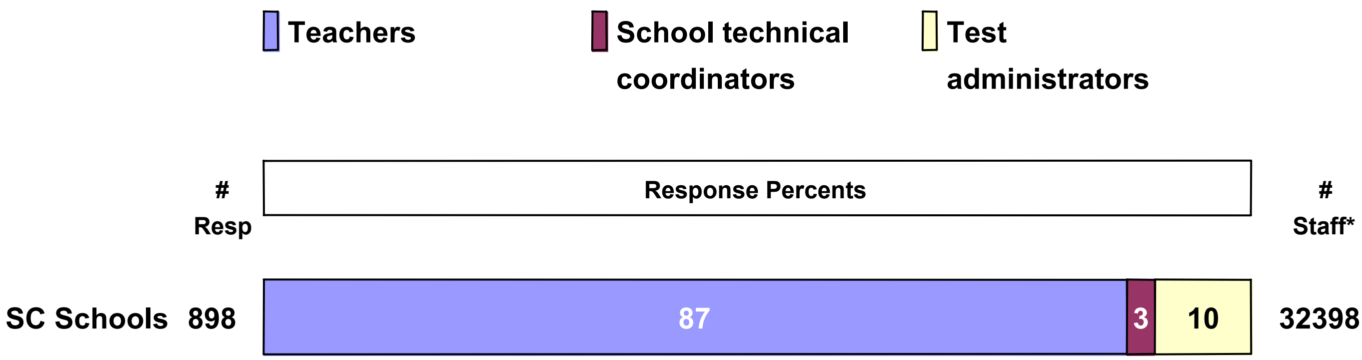
Please indicate the number of temporary personnel you needed during the 2005-2006 test administrations (e.g., monitors, volunteers, temps, etc).



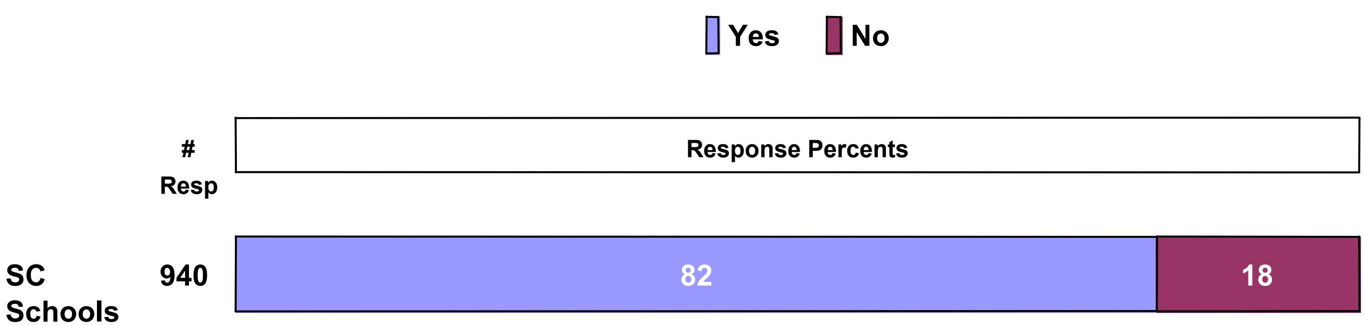
The mean for the State was 11.9.

School Test Administration Training

Please indicate the number of teachers/school technical coordinators/other test administrators you currently train annually for test administration.



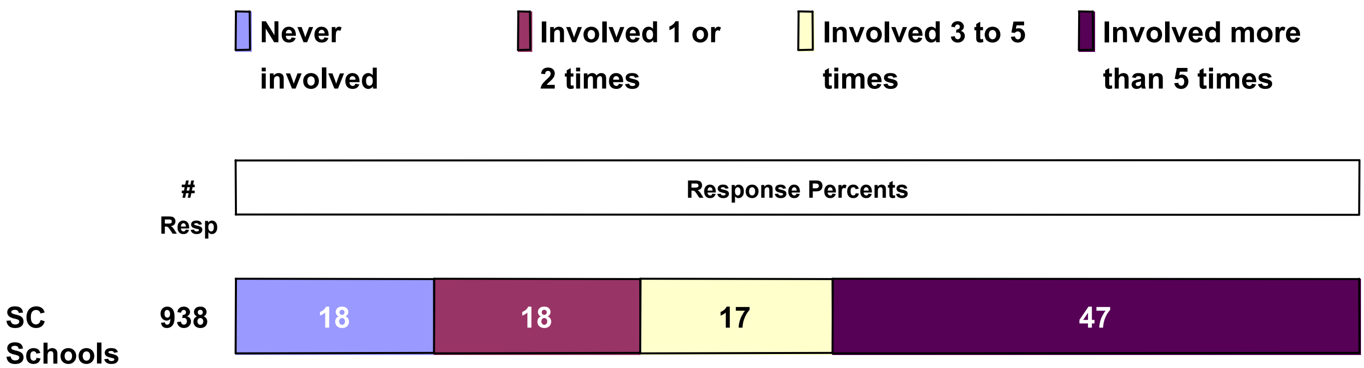
Is your current school involved in computerized testing?



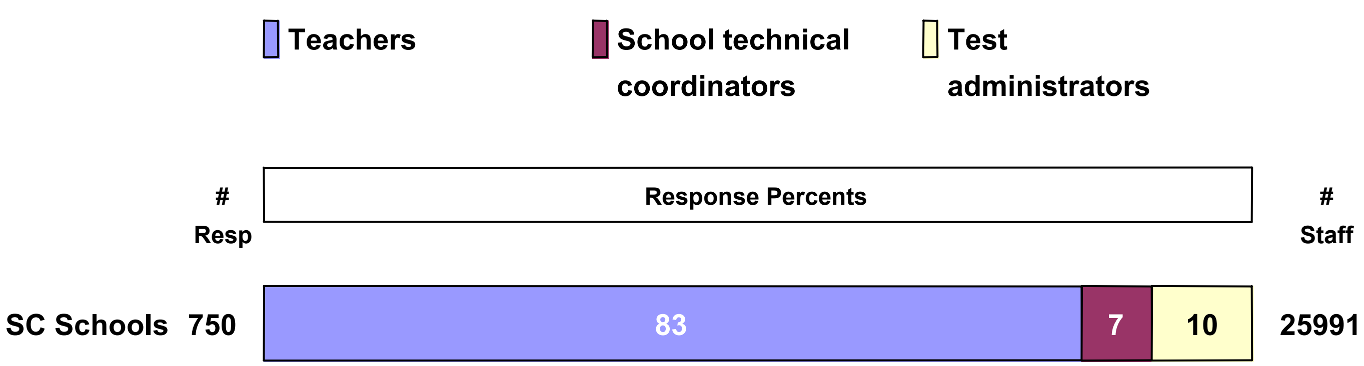
* The total number of staff reflects numbers reported by survey respondents.

School Test Administration Training

What is your level of experience with computerized testing?

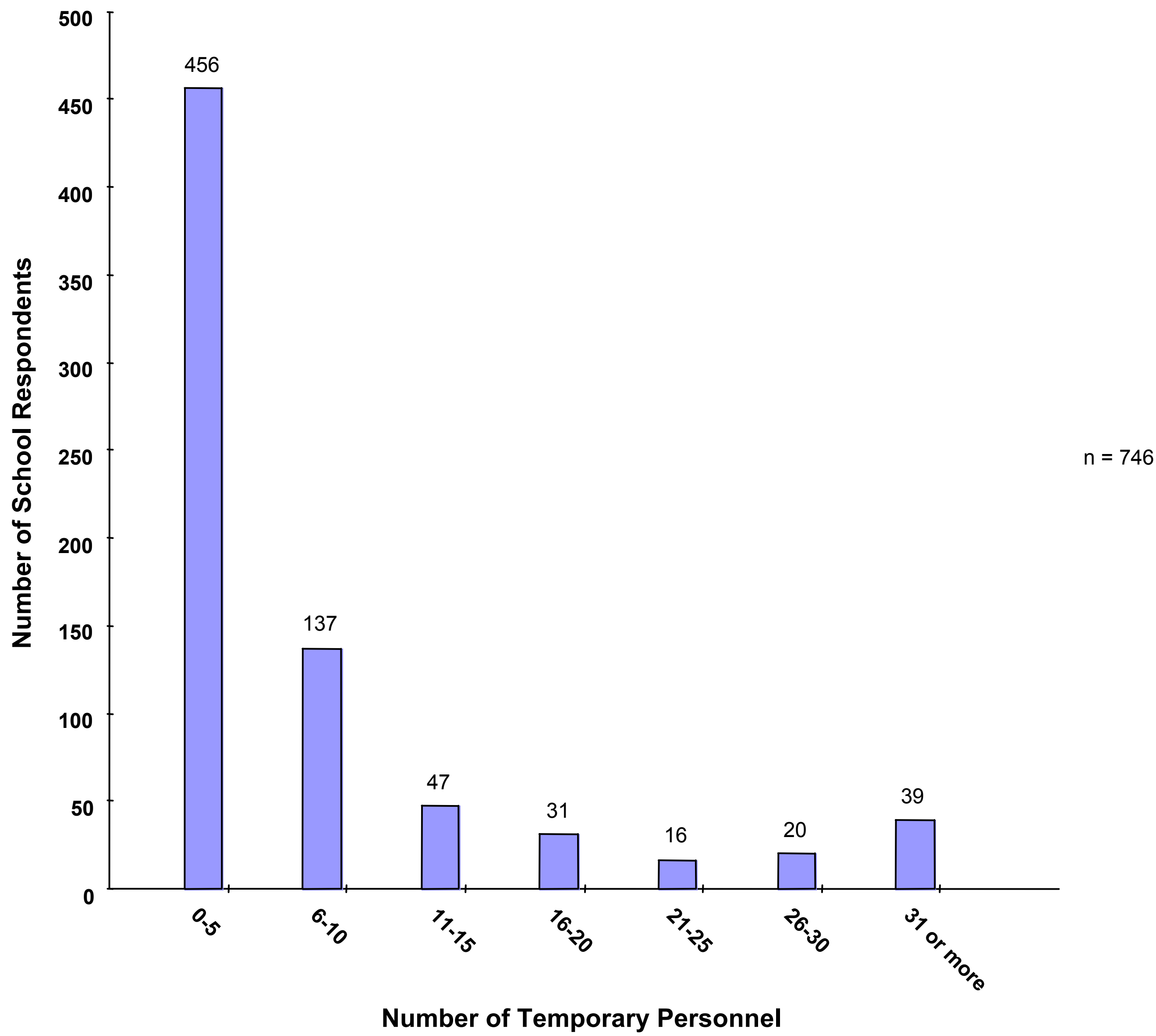


Please estimate the number of teachers/school technical coordinators/other test administrators you would need to train for test administration for a computerized test.



School Test Administration Training

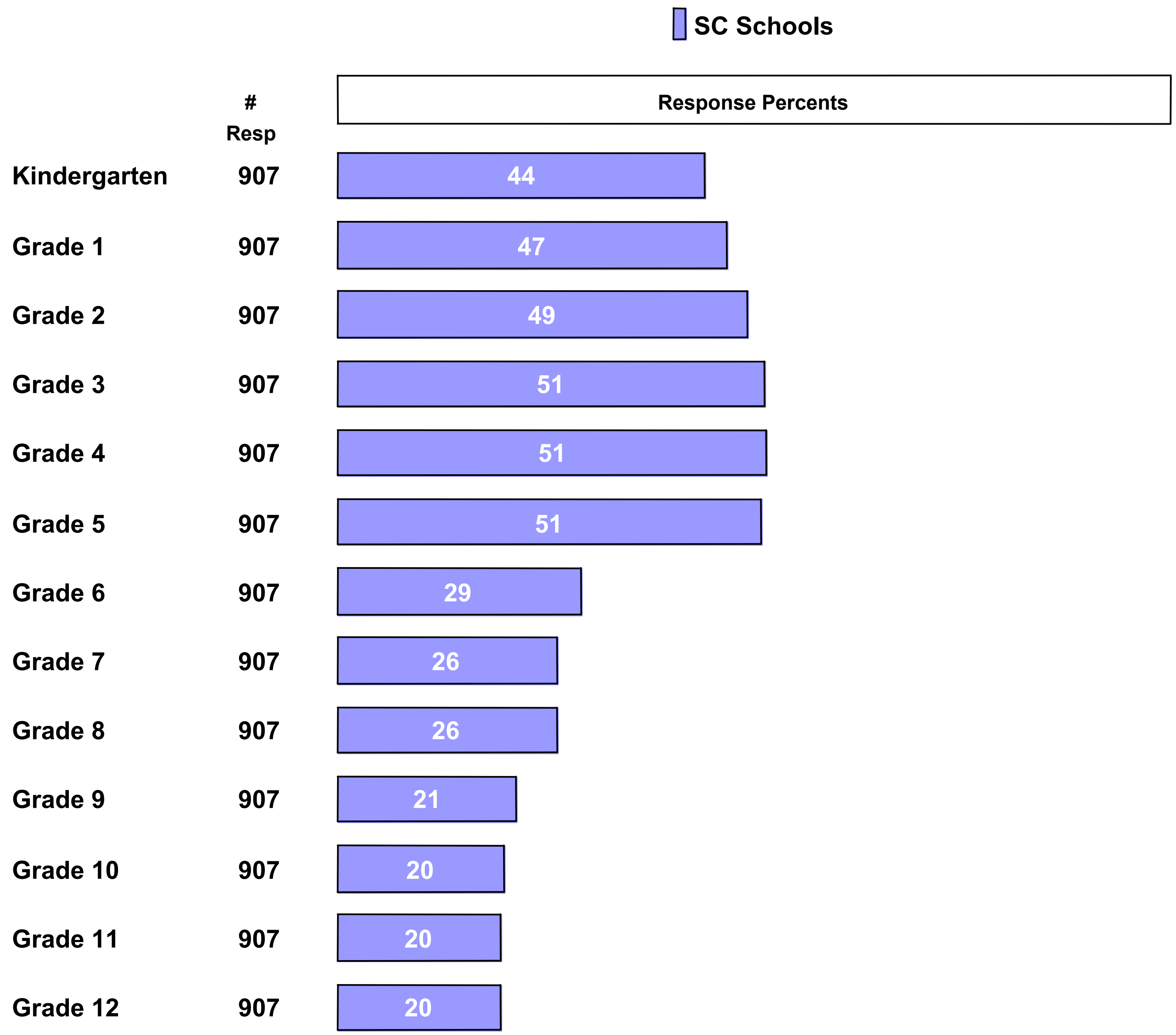
Please estimate the number of temporary personnel you would need during test administration for a computerized test.



The mean for the State was 8.6.

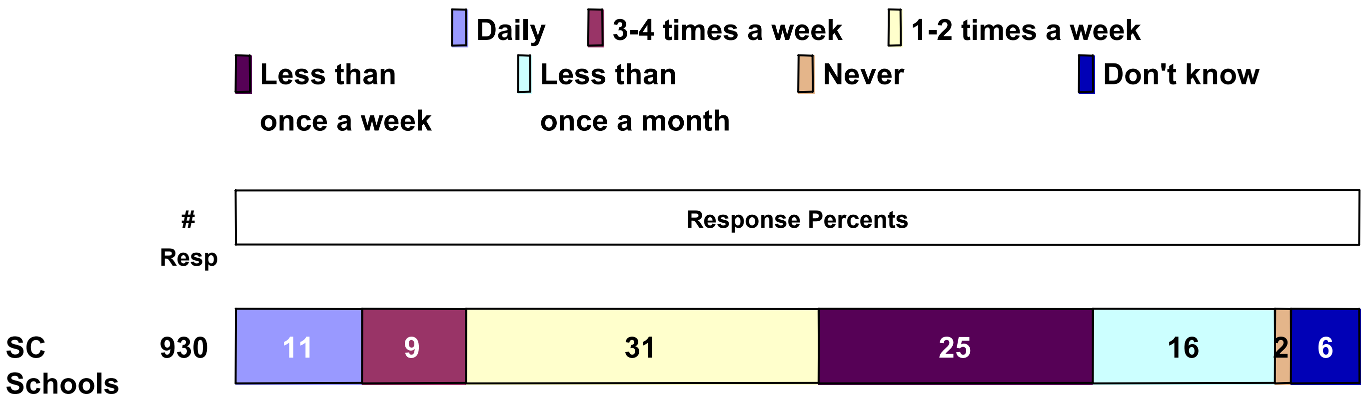
School Student Familiarity with Computers

At what grade levels are computers regularly used for classroom instruction at your school?

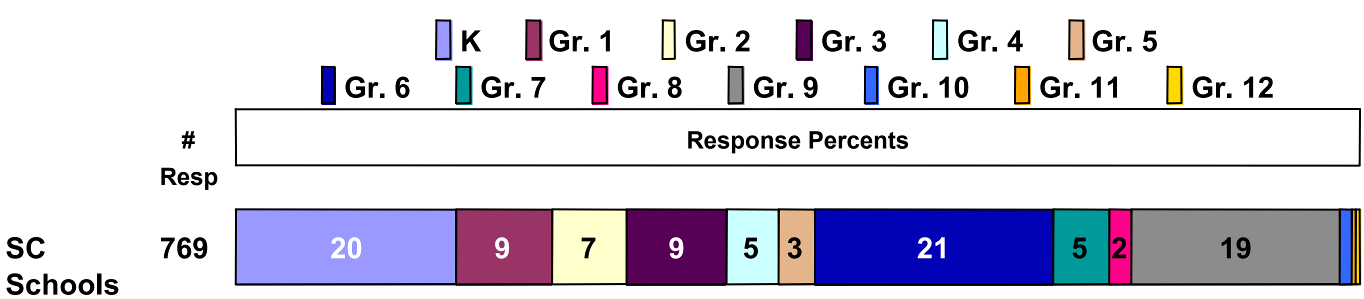


School Student Familiarity with Computers

How frequently do students at your school practice writing or composing on computers?

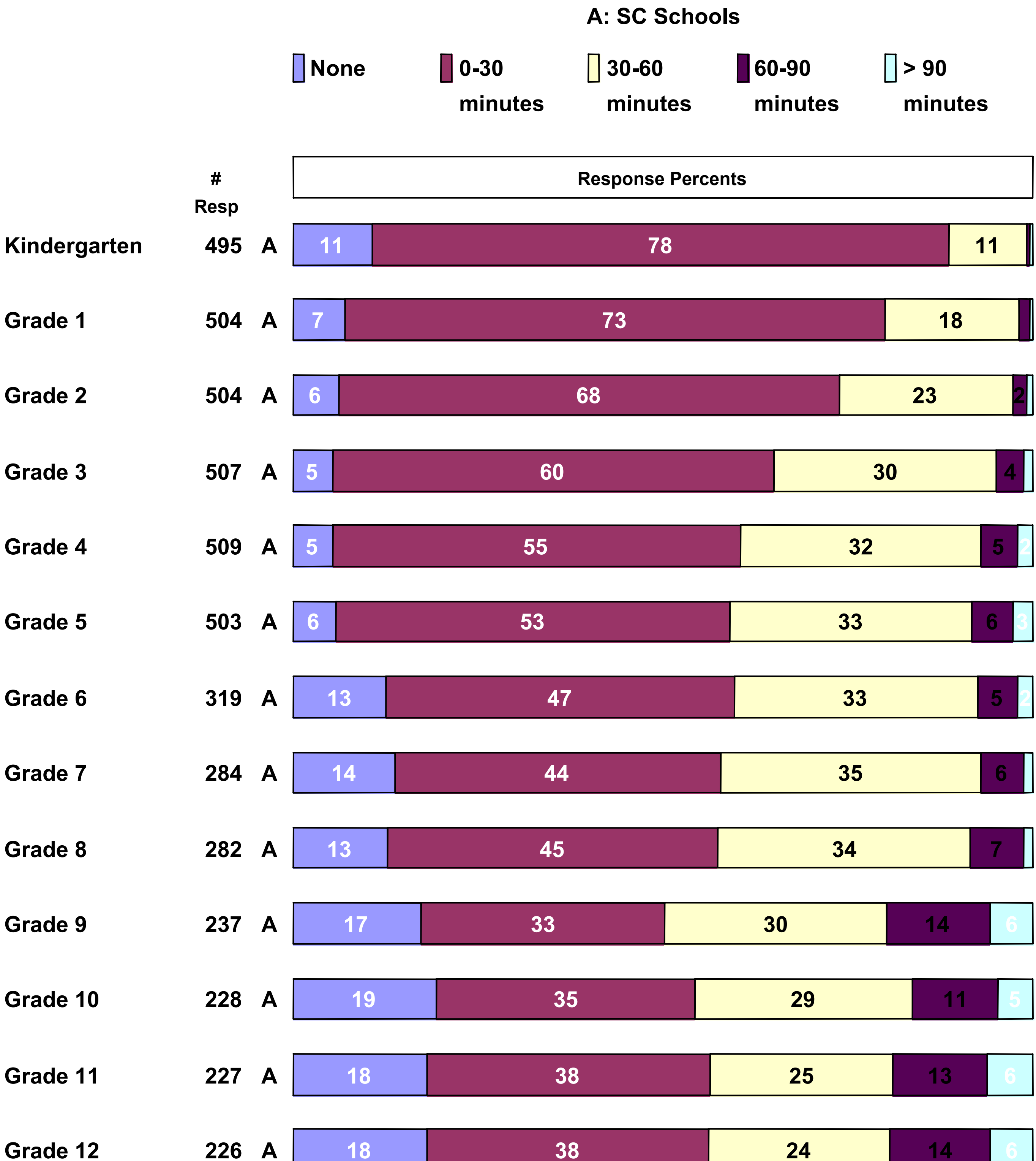


At your school, please indicate the lowest grade level at which students receive direct or formal instruction in keyboarding skills.



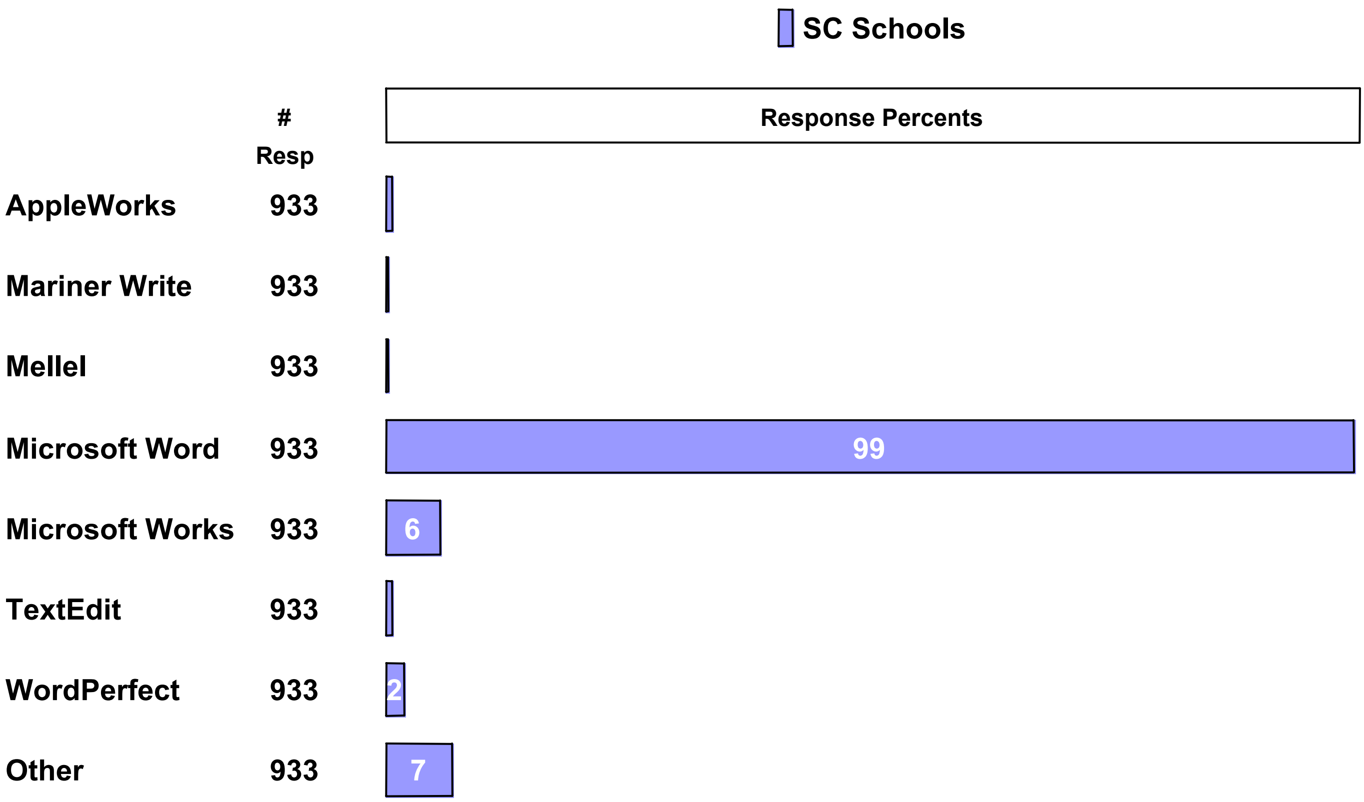
School Student Familiarity with Computers

At each grade level at your school, please estimate the amount of time students spend on a computer during a typical school day.



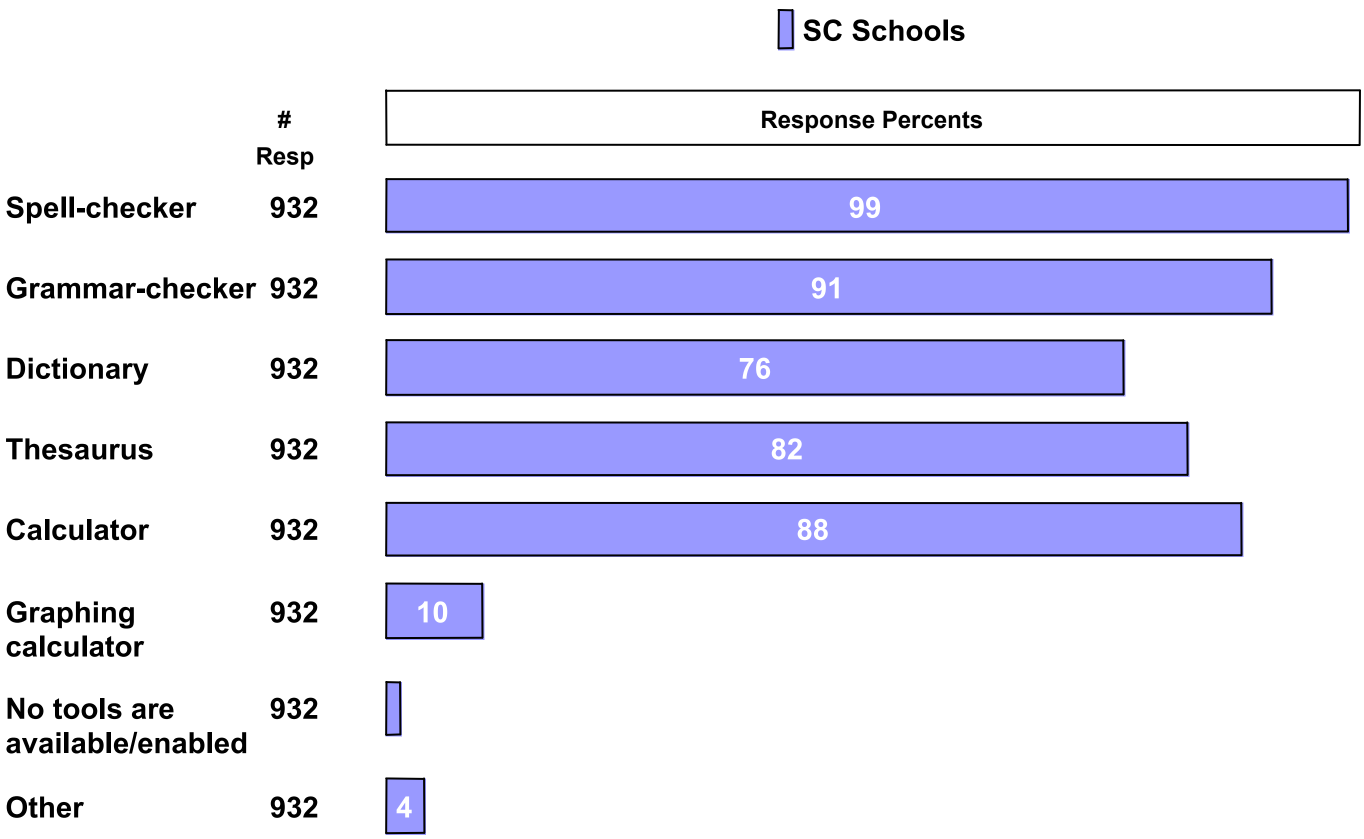
School Student Familiarity with Computers

What type(s) of word-processing software are used by students in their writing/composing on the computer?



School Student Familiarity with Computers

What tools are available to the students on the computers?



Study on the Feasibility and Cost of Converting the State Assessment Program to a Computer-Based or Computer-Adaptive Format

State Survey Comment Theme Report

STATE COMMENT THEME REPORT

One of the sources of information for the South Carolina Feasibility Study was a Web survey for the districts and schools within the State. The Web survey was used to collect current information about a variety of technology issues within districts and schools in South Carolina that are critical to assessing the feasibility of, and planning for, conversion to computer-based or computer-adaptive testing.

There were six (6) possible opportunities to provide comments on the SC Feasibility Web Survey; three comments at the school level and three at the district level. The following steps were taken to create themes for organizing and coding the comments:

1. During the administration period of the Web survey, a sample of comments was downloaded for comment code creation (79 comments).
2. Next, the comments were imported into the SPSS Text Analysis for Surveys 2.0 program.
3. Using the Extract Terms and Create Categories features of the software, initial comment themes were created. Parallel questions 68 and 84, 69 and 85, and 70 and 86 share the same comment themes.
4. The comment themes were validated with the complete set of comments and refined where necessary.
5. The comment themes were reviewed by the DRC and SC Teams and the final themes were applied to all the comments.
6. Finally, each comment was read through to manually validate the comment themes assigned to it.

Comment Themes

Table 1 lists the Web survey question text for the six comment opportunities.

Table 1: Web Survey Open-ended Questions

School Level	
68.	Thinking about all the needs of your school, please indicate the two or three biggest barriers to implementing computerized testing at your school.
69.	Thinking about all the needs of your school, please discuss the two or three biggest advantages to implementing computerized testing at your school.
70.	Do you have any other comments about how your school could best transition to computerized testing?
District Level	
84.	Thinking about all the needs of your district, please indicate the two or three biggest barriers to implementing computerized testing.
85.	Thinking about all the needs of your district, please discuss the two or three biggest advantages to implementing computerized testing.
86.	Do you have any other comments about how your district could best transition to computerized testing?

Tables 2, 3, and 4 list the comment themes and key words for each of the comment opportunities. The key words aid the analysis program and human coder in applying the correct theme to a comment.

Table 2: Comment Theme and Key Words for Questions 68 and 84 (Thinking about all the needs of your school or district, please indicate the two or three biggest barriers to implementing computerized testing.)

Theme	Key Words
Effect on Students	Computer skills and abilities, computer access, elementary, high school, student placement
Personnel	Training, hiring, technical personnel, teachers, staff, technology staff
Cost	Funding, money, purchase, upgrade, buy, printing, funds
Time	Testing window, testing period, time for preparation, scoring time, timeframe
Space	Location, classrooms, school labs, rooms, facility, computer labs
Technology	Hardware, computers, internet access, bandwidth, software, networks
Test Security	Security concerns
Implementation/ Transition	Implementation time, barrier to implementation, transition
Special Needs	Students with special requirements for testing
Miscellaneous	

Table 3: Comment Theme and Key Words for Questions 69 and 85 (Thinking about all the needs of your school or district, please discuss the two or three biggest advantages to implementing computerized testing.)

Theme	Key Words
Effect on Students	Computer skills and abilities, computer access, elementary, high school, student placement
Personnel	Training, hiring, technical personnel, teachers, staff, technology staff
Cost	Funding, money, purchase, upgrade, buy, printing, funds
Time	Testing window, testing period, time for preparation, scoring time, timeframe
Space	Location, classrooms, school labs, rooms, facility, computer labs
Technology	Hardware, computers, internet access, bandwidth, software, networks
Test Security	Security concerns
Current Testing Programs	MAP, PACT
Electronic Testing (eTesting) Process	Paper, shipping, security, materials, format, packing
Feedback	Scores, results, turnaround, data, diagnostic information
Special Needs	Students with special requirements for testing
Miscellaneous	

Table 4: Comment Theme and Key Words for Questions 70 and 86
(Do you have any other comments about how your school or district could best transition to computerized testing?)

Theme	Key Words
Effect on Students	Computer skills and abilities, computer access, elementary, high school, student placement
Personnel	Training, hiring, technical personnel, teachers, staff, technology staff
Cost	Funding, money, purchase, upgrade, buy, printing, funds
Time	Testing window, testing period, time for preparation, scoring time, timeframe
Space	Location, classrooms, school labs, rooms, facility, computer labs
Technology	Hardware, computers, internet access, bandwidth, software, networks
Implementation/ Transition	Implementation time, barrier to implementation, transition
Current Testing Programs	MAP, PACT
Special Needs	Students with special requirements for testing
Miscellaneous	

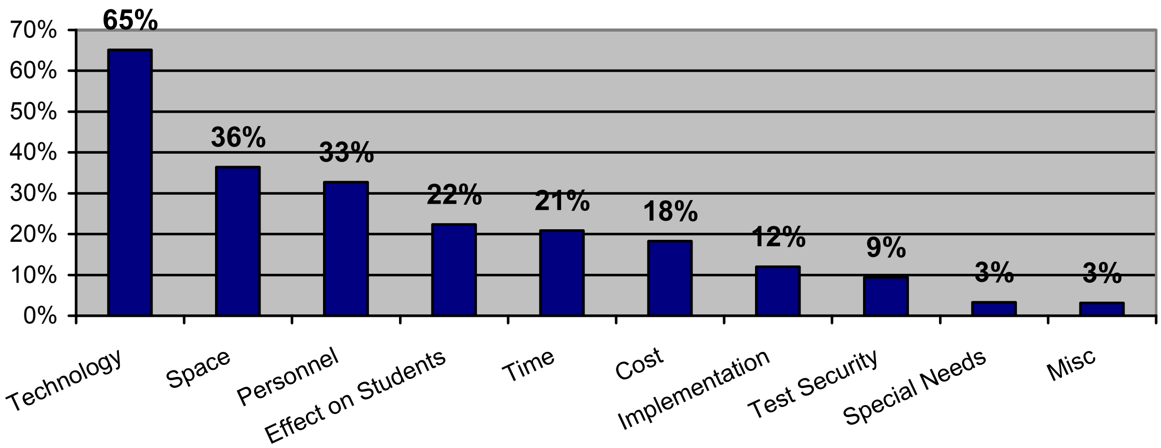
Results

There were 1,458 schools and districts invited to complete a Web survey. Of this total population, 948 schools and 86 districts returned surveys (1,034 total). Almost all of these returned surveys provided at least one comment (975 comments, 94% comment response). More specifically, 83 of the 86 district surveys and 892 of the 948 school surveys provided comments.

Question 68

There were 885 comments provided in response to Question 68: Thinking about all the needs of your school, please indicate the two or three biggest barriers to implementing computerized testing at your school. Figure One displays the percentage of comments containing each comment theme in descending order from most to least frequent. Note: The numbers do not add up to 100% because a comment could be assigned up to three comment themes. For example, if a respondent wrote about the amount of funding their school needs, the implications for students with special needs, and had concerns about test security, then their comment would be classified under the following themes: Cost, Special Needs, and Test Security.

Figure 1: Percent of Comments Containing Theme - Question 68 (Thinking about all the needs of your school, please indicate the two or three biggest barriers to implementing computerized testing at your school.)



The three most prevalent themes present in the comments for Question 68 (Thinking about all the needs of your school, please indicate the two or three biggest barriers to implementing computerized testing at your school) were Technology, Space, and Personnel. Within the Technology theme, concerns appear to center on the degree of technological readiness for testing, including issues of whether schools needed to upgrade their technology in order to be prepared for testing; whether there was an adequate number of computers for testing and whether there would be sufficient bandwidth for testing. Within the Space theme, concerns center on issues of having or managing testing space; including number or size of computer labs and whether additional labs could be built. Within the Personnel theme, concerns focus on issues of number, experience and type of personnel needed; including available technology staff, the familiarity of teachers with computerized testing, and additional test monitors needed.

Illustrative comments – Question 68

Many of the comments were thoughtful and instructive. Comments often enhance the information value of the quantitative survey results. Comments enable respondents to provide input that is not specifically covered in the other (closed-ended) survey questions and to provide their own school perspective. Following each theme listed below are sample comments (from all the comments submitted) that illustrate the themes. Note: The comments are presented as typed by the respondents.

TECHNOLOGY

We do not have enough centrally located computers to meet the needs of all students. Our existing computers are slow and unable to operate new software programs.

The two biggest barriers to implementing computerized testing at our school are: having enough computers available for students to complete the tests in a timely manner; and the amount of instruction students will not receive from the Keyboarding, PLATO and SuccessMaker Labs, due to relocating these classes from the lab sites.

Lack of computers. Almost all of our computers are 4 years old and we need to upgrade memory, etc. Class interruptions.

Students click answers before reading passages. Not enough up to date computers in the school. Some students have limited experiences with computers only at school.

The two barriers would be making sure that all computers are working and that technical support would be provided as needed to ensure that all students are tested.

SPACE

The two biggest barriers are the appropriate environmental space and enough time in the day to test students on the variety of tests that they would need to be tested on. We currently take the MAP test and this takes nine weeks a year to test k-5 students on three tests. There is too much demand on the current computer resources we have in place. The classroom setting is not appropriate to test children.

1. Lack of space to house the necessary number of computers needed to complete testing in a timely manner. 2. Lack of sufficient computers to complete testing in a timely manner. 3. Scheduling testing as to minimize disruptions to the regular educational schedule.

-Lack of space for additional computer lab (no classroom space available without additional building). -Lack of infrastructure to support additional computers in the building (need fiber optic wiring or wireless infrastructure). -Lack of personnel to run additional computer lab.

The need for more space and more computers.

1. Lack of technology funding 2. Not enough computer labs to make testing feasible 3. Lack of space to house additional labs 4. Too much time needed to complete testing.

PERSONNEL

My school does not have enough computer labs nor personnel to administer computerized testing.

Lack of experience on the part of administrators, lack of technical support at the school level, time it will take to complete the testing, computers and space for them.

1. A technician would [be] housed on site during all testing days. 2. Computer literate monitors. 3. Availability of computers for testing.

1. Security Concerns 2. No internet connectivity for students 3. Only one IT person for the entire district which is statewide.

Teachers not feeling computer iterate enough to handle computerized testing. Maintaining enough up-to-date computers from year to year for testing.

EFFECT ON STUDENTS

Our 4th and 5th graders are not taught keyboarding. This makes publishing their writing very slow.

Computer malfunctions while a test is in progress would be distracting and an impede a student's progress.

Computer Accessibility Space for testing Familiarity with computers by students

1) Training of staff 2) Sufficient Hardware (computers) 3) Variance in computer literacy among students

The age level and computer experience of our children. Our large student enrollment and the impact on instructional time. Lack of personnel to adequately monitor such a test.

TIME

Length of time due to size of school

1. Getting all the students tested in the given timeframe. 2. The students may not take another computerized test seriously.

Possible time restraints, number of available computers which could include financial issues and space issues, large number of test accommodations.

-Children not talking it as serious as paper/pencil test -Time it will take to cycle students through lab -Time computer lab will be closed for student instruction during testing weeks -Writing portion could take much longer for students to complete

Time to schedule all students in for all subtests; Spreads the testing window -Possibly impacting in test apathy among students.

Cost

Cost of hardware Cost of software

Funding for switches, cabling, computers, district support technicians, and building money for constructing additional classrooms to house computer labs.

Monies needed to maintain computers and computer system and limited number of computers available in a proctored setting for Test use and the amount of time necessary to complete all Testing.

1. Training teachers to administer the tests. 2. Funds to upgrade hardware to run the tests. 3. Students may just "push the key" for an answer rather than read through a reading passage or work out a math problem.

- Lack of funding that would be necessary to implement computerized testing for state mandated assessments.   - Lack of computers, resources, and tech support that would be required to manage and maintain the number of computers that would be required to implement computerized testing for state mandated assessments.

IMPLEMENTATION/TRANSITION

The scheduling of the students would be difficult to allow access to the computers in the computer lab. (300 students-53 computers-4 subjects). Would the students all take the test in the morning or would they take it all day? Would all students be able to perform at their best at different times of the day? How would accommodations be allowed for small groups and oral testing?

The transition from paper to computer would be the biggest barrier (the mindset of the students).

The three biggest barriers to implementing computerized testing at my school are: 1. The amount of time this will take. 2. The number of computers we have at present that 3. How will it be implemented?

1. Technical problems 2. Transitioning from paper to the computer screen 3. Fatigue, zoning out

A barrier for implementation would be scheduling.

TEST SECURITY

Computer lab scheduling, security issues, test monitoring, reliability of scores should technical problems arise.

test security between administrations, too few computers in our building, testing late in the day to accommodate all students, server issues

1. Test security. 2. Validity of using keyboard vs. paper/pencil 3. Lab space

Only 2 centralized locations for testing. Inadequate number of computers. Not sure school server will accommodate needs of centralized testing. Testing security could be compromised.

More computer labs. Faster internet connection. Security.

SPECIAL NEEDS

Having enough time to test all students. Implementing modifications of special needs students for testing.

Special Needs Accommodations

1. We have a high percentage of special needs students in our population that would require additional assistance. 2. Our equipment is outdated. 3. We have a high degree of student absenteeism. 4. State infrastructure would have to be upgraded

We have one computer lab to accommodate a large population of students. This would present problems with test item security, small group testing, and testing for special needs students.

Since we have 3 LD Self-Contained classes, we have a large number of students with test modifications and accommodations. The computerized test would cause us to have more test administrators to individual and small group testing.

MISCELLANEOUS

I do not see any barriers. We have been doing computerized testing of the NWEA Map test for 3 years and it has been a very positive experience.

There are currently no barriers to test implementation. We could begin CBT with a month's notice.

We don't have any.

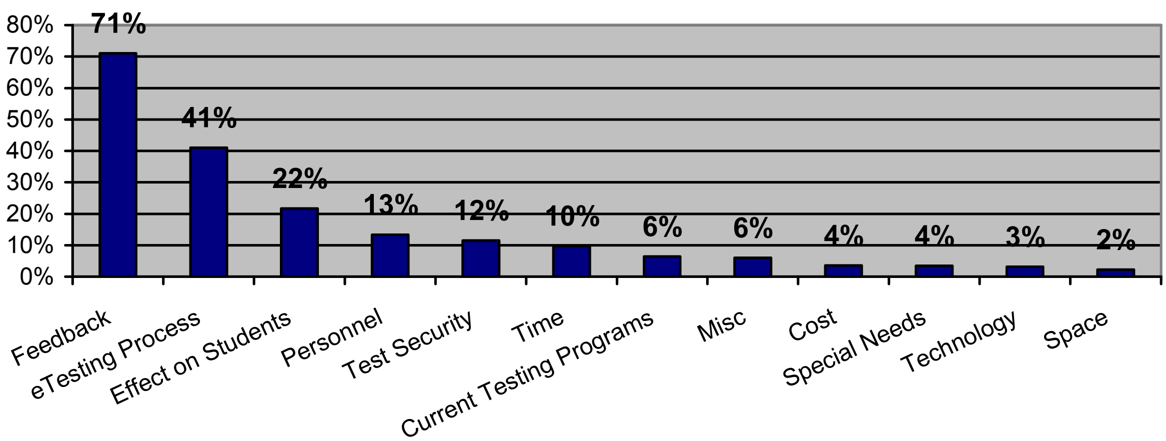
None. We want to do it.

We currently have computerized testing at our school. Assuming that a suitable product is selected, there are no barriers.

Question 69

There were 863 comments provided in response to Question 69: Thinking about all the needs of your school, please discuss the two or three biggest advantages to implementing computerized testing at your school. Figure Two displays the percentage of comments containing each comment theme in descending order from most to least frequent. Note: The numbers do not add up to 100% because a comment could be assigned up to three comment themes.

Figure 2: Percent of Comments Containing Theme – Question 69 (Thinking about all the needs of your school, please discuss the two or three biggest advantages to implementing computerized testing at your school.)



The three most prevalent themes present in the comments for Question 69 (Thinking about all the needs of your school, please discuss the two or three biggest advantages to implementing computerized testing at your school.) were Feedback, Electronic Testing (eTesting) Process, and Effect on Students. Within the Feedback theme, comments appear to center on the speed of receiving scores and the advantages of using feedback for planning. Within the Electronic Testing (eTesting) Process theme, comments center on the advantages of computerized testing versus paper and pencil testing; including streamlined processes, easier material collection, and less paperwork.

Within the Effect on Students theme, comments focus on students' positive perceptions of and learning abilities with computers.

Illustrative comments – Question 69

Many of the comments were thoughtful and instructive. Comments often enhance the information value of the quantitative survey results. Comments enable respondents to provide input that is not specifically covered in the other (closed-ended) survey questions and to provide their own school perspective. Following each theme listed below are sample comments (from all the comments submitted) that illustrate the themes. Note: The comments are presented as typed by the respondents.

FEEDBACK

> The advantage would be receiving scores in a timely manner. Test security would be less of an ordeal. If it is adaptive computerized testing, then oral test administrators would not be needed.

> The biggest advantages to implementing computerized testing is that it saves on having to deal with the numerous books (counting, sorting, accountability, etc.). Another advantage is that the scores come back much quicker.

> Three of the biggest advantages to implementing computerized testing at our school are: students like to work on computers better than taking paper and pencil tests; computerized testing should help with the elimination of time used for preparation and handling of test booklets; and we should receive test results faster than the current wait time for receiving test results.

FEEDBACK (CONTINUED)

Alternative way to test students. Provides students and teachers with timely feedback (MAP is an example) as opposed to utilizing PACT which provides little to no feedback in a non-friendly time frame.

We LOVE Measures of Academic Progress, and especially like that we can actually do something with the results, rather than having to wait until the next year to find out how students did.

ELECTRONIC TESTING (ETESTING) PROCESS

Computerized test will simplify the process!!

Obviously, the testing process would be more streamlined and simplified.

We feel we have the flexibility among the staff and students to make it work. Our positive attitudes towards computers and the realization that it is time to change how we test.

Hopefully, quicker information on students' test results. Paperless!!!! Technology is the way to go. We need to take advantage of it!

There would not be as much paperwork or as many security issues.

EFFECT ON STUDENTS

Some students perform better using this mode.

Students love working on computers

I feel that it would save a lot of paper. My students are computer geniuses and would probably perform better if the test was computerized. My students are used to taking assessments via the computer and I think that for students with weaker ability levels, this would assist them in feeling more comfortable and doing better.

Flexibility of scheduling. Convenience for students.

Student interest, faster results, less chance of paper errors, security.

Personnel

Less paper work and man hours for the test coordinators. Easier to administer these tests. More specific and timely results.

1. Fewer people would be involved in the testing process. 2. Results of the tests would return in a shorter period of time. 3. Time packing materials for shipping.

We feel we have the flexibility among the staff and students to make it work. Our positive attitudes towards computers and the realization that it is time to change how we test.

1. Alleviate many personnel from testing duties- Would allow for more time for direct instruction 2. Faster results 3. More kid-friendly 4. Less cumbersome testing material handling

Testing environment would be the same for all students. Legibility problems eliminated. Guidance counselor would have more time for guidance rather than test monitoring and preparation.

TEST SECURITY

The biggest advantages would be security and speed of returned assessment results.

Time packing, preparing test materials would be saved. Test security would be better in some respects- losing materials.

1. Immediate results 2. Better system of security 3. Possibility of maintaining better student records to observe growth patterns 4. Would free up guidance counselors 5. Cost efficient

Eliminates the book and security issue. We MAP test online now. Saves time.

More effective and efficient Less time to prepare for test administration Security is automatic (a very big headache for schools with storage, passing out, etc of paper booklet tests) Quick Results!

TIME

Shorter time frame for getting test results Much less paper would be needed to practice and test Rules of security would not have to be as strict because less personnel need to see and administer tests Students would stay focused on computerized testing more than the rigorous paper and pencil tests.

Time to take test, immediate feedback.

Three advantages of computerized testing at our school are: 1) Less paper work 2) Security issue   3) Time saver

Eliminates the book and security issue. We MAP test online now. Saves time.

Time, convenience, security.

CURRENT TESTING PROGRAMS

Implementing a NWEA MAP like test would be advantageous since it can be done in a lot of setting[s] simultaneously. Being able to get results in a more timely basis.

Students practice using EduTest or MAP. The skills may carry over better to computerized testing for PACT

Through computerized MAP testing we can get test scores quickly and chart individual student progress 3 times throughout the year.

Students are comfortable/familiar with computers since we already do computerized MAP testing. Schools would have results faster than paper-based tests.

We really like the MAP test for its quick results. It can be used diagnostically unlike PACT. It SHOULD eliminate much of the paperwork, packing tests for shipment, counting and re-counting test booklets, etc. that happens with paper tests.

MISCELLANEOUS

I can think of no advantages that would outweigh the disadvantages and lack of time and resources to implement computerized testing.

Need more information to address this question

Advantages to computerized administration of PACT at the school level - NONE

There are no advantages.

None

COST

Time saver for test administrator Faster results and feedback for use in instruction Cost benefit for the state

receive results faster, decreased cost of printing tests, less chance of testing being misplaced or mishandled

1. Getting results faster 2. Saving money

1. Immediate results 2. Better system of security 3. Possibility of maintaining better student records to observe growth patterns 4. Would free up guidance counselors 5. Cost efficient

No direct advantage to the school exactly...I'm sure it would save the state massive amounts of money which would trickle down to the schools in the state eventually(hopefully).

SPECIAL NEEDS

Computerized testing would allow us to test students with special needs at various settings. It would also cut down on the need for transporting paper forms to and from 9 different schools.

1. Special needs students will be able to achieve at a higher rate through computer testing. 2. Our students are currently participating in MAP testing, and are familiar with taking tests on the computer.

1. Instant results/feedback 2. It may address different styles of testing.

Less paper work. Less chance of losing/misplacing materials. Help students with disabilities. Get results back quicker

1. Immediate feedback. 2. Save money for postage. 3. IEPs would be easier to accommodate.

TECHNOLOGY

Forcing replacement of old equipment Forcing Bandwidth

We would get our equipment upgraded and we would get to add computers to our existing fleet.

The biggest advantages would be the relatively new computers and a computer lab located on each grade level pod.

I believe that the biggest advantages that our school has when it comes to the implementation of computerized testing are: Up to date Computers, Instant Scores, Student placement for the upcoming year.

1. Results should be available to teachers sooner. 2. Possible upgrades to computers in the school.

SPACE

Accuracy and speed, but only if there is more technology and space allocations!

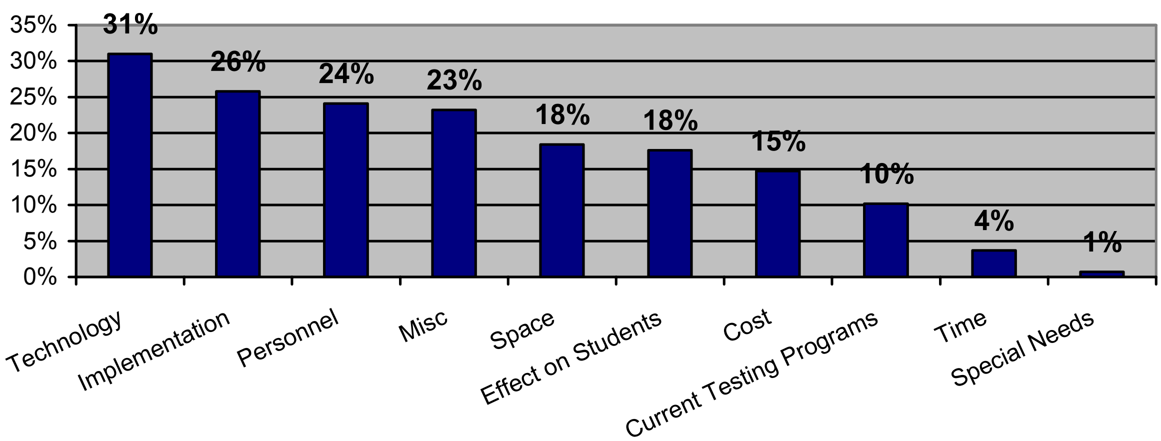
Less time testing, less money to test and less space to test are the advantages

There is plenty of control space and teachers can watch effectively as learners are testing.

Question 70

There were 539 comments provided in response to Question 70: Do you have any other comments about how your school could best transition to computerized testing? Figure Three displays the percentage of comments containing each comment theme in descending order from most to least frequent. Note: The numbers do not add up to 100% because a comment could be assigned up to three comment themes.

Figure 3: Percent of Comments Containing Theme – Question 70 (Do you have any other comments about how your school could best transition to computerized testing?)



The three most prevalent themes present in the comments for Question 70 (Do you have any other comments about how your school could best transition to computerized testing?) were Technology, Implementation/Transition, and Personnel. Within the Technology theme, concerns appear to center on the degree of technological readiness for testing, including issues of whether schools needed to upgrade their technology in order to be prepared for testing; whether there was an adequate number of computers for testing and whether there would be sufficient bandwidth for testing. Within the Implementation/Transition theme, concerns center on issues of scheduling computerized testing versus paper and pencil testing and pilot testing. Within the Personnel theme, concerns focus on issues of number, experience and type of personnel needed; including available technology staff, the familiarity of teachers with computerized testing, and additional test monitors needed.

Illustrative comments – Question 70

Many of the comments were thoughtful and instructive. Comments often enhance the information value of the quantitative survey results. Comments enable respondents to provide input that is not specifically covered in the other (closed-ended) survey questions and to provide their own school perspective. Following each theme listed below are sample comments (from all the comments submitted) that illustrate the themes. Note: The comments are presented as typed by the respondents.

TECHNOLOGY

> To be feasible, our school would have to have more computers

> District plan for computer maintenance and upgrade of existing computers.

> More computers and space is needed for this to actually work...funding needs to be available for this to occur.

> We desperately need more funding for technology to include providing more computers and the staff necessary to operate a computer lab and to teach keyboarding to students.

> Proper training for personnel and proper equipment for testing.

IMPLEMENTATION/TRANSITION

Introduce the computerized testing as a mock trial testing before implementing. Implement one grade per year to see if this is beneficial.

A transition could take place, one grade at a time, after it has been piloted and all the kinks have been worked out elsewhere. I would not want our school to be one of the pilot schools.

In order to transition to computerized testing, we need time to insure that our students have been taught adequate Keyboarding and word processing skills. Hopefully, this concept of computerized testing will start as a pilot project.

We are presently using computer based testing so the transition in my opinion would be smooth. One problem with transitioning will be to construct a schedule that ensures that every student in the school is tested in a timely manner.

Because we have been involved with MAPS testing for the past 3-4 years, I do not feel like the transition would be difficult for our school. I think teachers and students would welcome the change.

PERSONNEL

We would need a consistent plan(from the district level) for keeping teachers trained - also need a district-wide plan for updating computers for the future.

Training of one teacher administrator and one or two monitors instead of training every teacher Possibly adding a mobile lab during testing Having a "practice" computerized test for all grades tested

Proper training for personnel and proper equipment for testing.

Make sure that the staff is well trained and they will feel comfortable with the new process. Also, we want to be sure that we have enough updated computers that could be used in order to cut down on computer testing time.

*full time technology person at each school *another computer lab *mobile labs

MISCELLANEOUS

Not in favor of computerized testing.

We are strongly in favor of CBT PACT administration.

No further comments.

I think computer testing would be great.

None

SPACE

We need more space in our building before we could accommodate this type of testing. Our software would also most likely need to be updated.

School would need space to accommodate a testing environment. Also, the school would need a dedicated budget for computer hardware and upgrades to labs and the network connections.

In order to transition, we need more computers and more space. Every child needs a computer in the classroom so that they can be assessed daily with computerized testing. The students must be comfortable with using the computers for testing.

There is a need to make sure schools have the bandwidth and computer resources to test kids in a timely manner. Also, cramped computer labs are not an ideal setting for testing, so space issues need to be addressed.

More computers and space is needed for this to actually work...funding needs to be available for this to occur.

EFFECT ON STUDENTS

In order to transition to computerized testing, we need time to insure that our students have been taught adequate Keyboarding and word processing skills. Hopefully, this concept of computerized testing will start as a pilot project.

Make available computerized practices for the students to use throughout the year. Set up more labs or ITEC classrooms in the schools.

Our students will transition nicely to computerized testing because they are familiar with MAP testing and working on computer based programs daily such as: Compass Learning, Renaissance Place, Destiny, and various other programs available at the school.

Laptop computers are needed for every child along with an operating wireless network. Assistive technology for special needs students. Alternative Plan for students who would normally be using the computer labs when labs are used for testing.

It would be nice to have a computer for every student, but unless they were laptops, where would we put them?

COST

Equipment funding and physical space requirements are our biggest barriers to overcome to support the transition.

Additional funding by the state for the purpose of adding additional labs to the building would be the most help.

Money to upgrade and maintain would be beneficial and we would need someone school-based to assist with the maintenance. We would also need time and money to train faculty and staff.

We would need much funding for updated computer purchasing and additional space for setting up labs.

Because our school is in a rapidly growing area, planning for additional classroom space may be necessary. The state must not pass this cost on to the districts. The administration of accountability instruments should not interfere with instruction

CURRENT TESTING PROGRAMS

We currently use MAP testing in our district and I feel it is a far more valuable tool than the results of PACT. We are able to test all our students in grades 3-5 during a three-week period without disrupting their normal day. MAP is administered during the time they would normally be in the computer lab during the week.

Since our school is actively involved with MAP testing, students are very familiar with this format of testing. Therefore, the transition from paper testing to computerized testing would not be difficult for them.

We already do MAP testing, Academy of Reading, Success Maker and had Tungsten. We believe very little transition would be needed. We would be very interested in piloting a computerized PACT test.

I believe that our school is ready to transition to computerized testing. The students are familiar with STAR testing, Accelerated Reader, and MAP testing

I feel that the transition to computerized testing would be fairly easy for our school since we have been involved in MAP testing for several years.

TIME

Could we have more time to complete testing. For example, currently HSAP Ela testing is done in two days and math is done in one day. With so many students having to test in a lab, could there be more days added to accomplish the task at Stratford. Or maybe a lab of 100 computers could be set up.

We have the knowledge for computerized testing since we have been using MAP for years. Time and equipment would be areas of our concern.

Please provide adequate equipment, time and training.

SPECIAL NEEDS

Laptop computers are needed for every child along with an operating wireless network. Assistive technology for special needs students. Alternative Plan for students who would normally be using the computer labs when labs are used for testing.

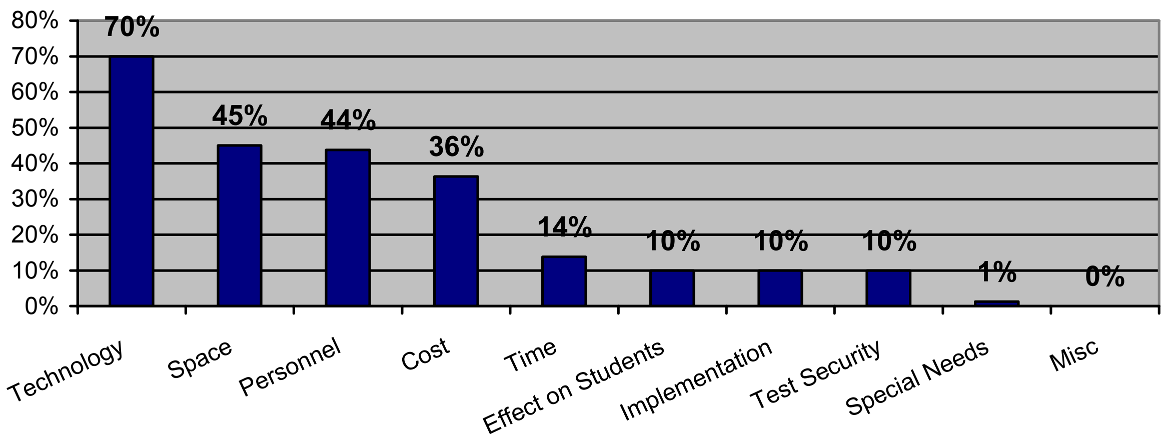
MOBILE LABS, KEYBOARDING CLASSES, ADDITIONAL COMPUTER LABS. WE DO NOT HAVE EARPHONES/HEADSETS OR OTHER DEVICES FOR SPECIAL NEEDS STUDENTS.

Need to have some dialogue about how special ed students will be accommodated before this is implemented to ensure IEP's are accurate.

Question 84

There were 80 comments provided in response to Question 84: Thinking about all the needs of your district, please indicate the two or three biggest barriers to implementing computerized testing. Figure Four displays the percentage of comments containing each comment theme in descending order from most to least frequent. Note: The numbers do not add up to 100% because a comment could be assigned up to three comment themes.

Figure 4: Percent of Comments Containing Theme - Question 84 (Thinking about all the needs of your district, please indicate the two or three biggest barriers to implementing computerized testing.)



The three most prevalent themes present in the comments for Question 84 (Thinking about all the needs of your district, please indicate the two or three biggest barriers to implementing computerized testing) were Technology, Space, and Personnel. This is parallel to the results for Question 68. Within the Technology theme, concerns appear to center on the degree of technological readiness for testing, including issues of whether schools needed to upgrade their technology in order to be prepared for testing; whether there was an adequate number of computers for testing and whether there would be sufficient bandwidth for testing. Within the Space theme, concerns center on issues of having or managing testing space; including number or size of computer labs and whether additional labs could be built. Within the Personnel theme, concerns focus on issues of number, experience and type of personnel needed; including available technology staff, the familiarity of teachers with computerized testing, and additional test monitors needed.

Illustrative comments – Question 84

Many of the comments were thoughtful and instructive. Comments often enhance the information value of the quantitative survey results. Comments enable respondents to provide input that is not specifically covered in the other (closed-ended) survey questions and to provide their own district perspective. Following each theme listed below are sample comments (from all the comments submitted) that illustrate the themes. Note: The comments are presented as typed by the respondents.

TECHNOLOGY

> 1 The need for more updated workstations 2 A testing environment conducive to testing. The computer labs are cramped.

> 1. Need more BAN width and funding to upgrade it, 2. Need more funding for network monitoring, management, software and staffing, 3. Need more funding for teacher training to implement computerized testing, 4. Need more time for data preparation, 5. Need communication from state at least 4 months prior to implementation for advanced planning.

> 1. Having enough computers to test all our students in an efficient timeframe. 2. The impact of testing on instruction and remediation 3. Bandwidth issues if the test is conducted over the Internet.

TECHNOLOGY (CONTINUED)

Insufficient number of computers in the schools. Inadequate budget to purchase/maintain number of computers necessary. Security issues.

Lack of computers, physical space for additional computer labs and technical staff signed to handle testing in large scale. Lack of funding allocation to maintain computers and acquire new advanced computers.

SPACE

Space to create a lab, computers for a testing lab

Brick and mortar (space)/funds, personnel/funds, computers/funds

The number of computers, the rooms to put them in, and the money to maintain the computers.

Physical space for testing and numbers of computers to complete testing in a timely manner.

District infrastructure (power, space) Number of computers Certified staff with inadequate technology skills

PERSONNEL

1. Resources: Computers and IT staff at schools 2. Training School Test Coordinators

The biggest barrier to implementing computerized testing is lack of funds. The next biggest barrier is lack of personnel to train teachers and students so that they will be technologically proficient. There is a definite need for a structured plan for technology proficiencies to be implemented in a graduated implementation model from 4K to 12th grade for students, and funding is needed to prepare teachers as well.

1. Space 2. Funds to equip additional labs. 3. Funds to hire personnel to manage lab.

Sufficient number of computers or laptops. Training personnel.

BANDWIDTH, computers that do not interfere with instruction, funding to support such and SUPPORT STAFF.

COST

There is not an annual budget/policy/plan for computer replacement in our district. There are not enough computers in every classroom for computerized testing. There are not enough computer labs in every school site for computerized testing. Current bandwidth is inadequate for computerized testing.

Funding for hardware, including wiring and electrical needs Funding for necessary personnel Physical space (classroom lab locations)

Time

The time involved in testing all of our students falling into the various categories. We cannot afford to lose instructional time by tying up what technology we have with another test. Currently we utilize MAP at least twice and with plans for three times a year and this is already a problem. Even if we had enough equipment, we don't have the manpower or facilities.

Money for computers, time to test, personnel

Effect on Students

The keyboarding skill level of the majority of students is not high enough to not cause frustration in grades 3-7. We do not have adequate numbers of computers in a lab environment available for testing purposes in most schools. Limited funds to purchase computers.

Funding Space for computer labs Reliability of results (student keyboarding skills)

Implementation/Transition

1. Number of computers available to accommodate class size. 2. Scheduling to coordinate testing of large numbers of students.

1. The length of the PACT test and the necessity to expand the test window throughout the school day instead of limiting it to the morning. 2. The length of the testing window would have to be longer to accommodate testing with the existing machines. 3. The small number of technicians to deal with equipment problems may be problematic.

TEST SECURITY

1. Security concerns. 2. No internet connectivity for students. 3. Only one IT personnel for the entire state-wide district.

Security, Internet bandwidth, number of workstations

SPECIAL NEEDS

1.Standardized workstation hardware and sufficient bandwidth with standard specifications and necessary software for all testing scenarios, including, but not limited to, special education specialized software and high school specific software. 2. Technology support personnel at each school during all designated testing dates for immediate response to computer troubleshooting for immediate assistance. 3. Necessary classroom space with all necessary testing hardware configured for security (i.e. privacy screens) on monitors, uninterruptible power supplies and software as not to displace existing students who may be using computer labs for other instructional purposes

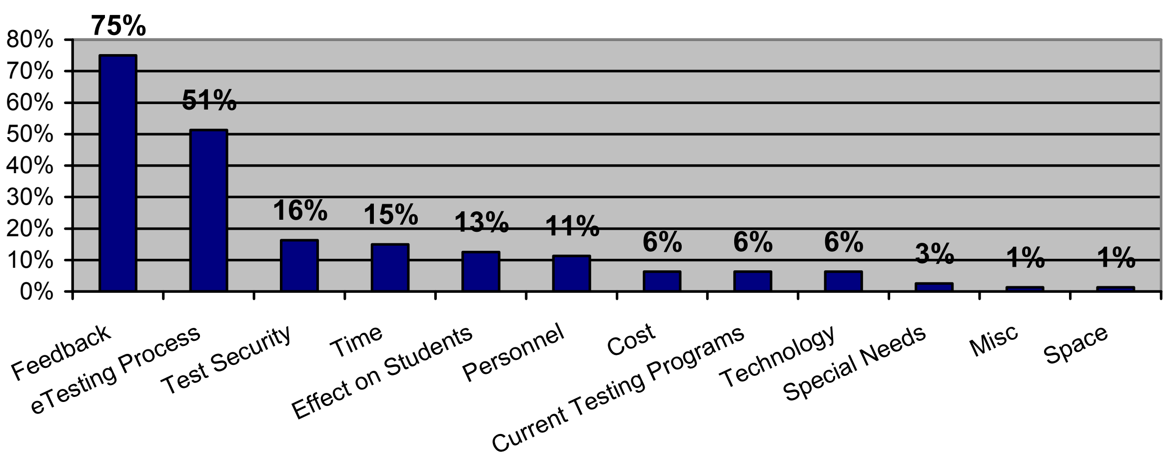
MISCELLANEOUS

No comments provided in this category.

Question 85

There were 80 comments provided in response to Question 85: Thinking about all the needs of your district, please discuss the two or three biggest advantages to implementing computerized testing. Figure Five displays the percentage of comments containing each comment theme in descending order from most to least frequent. Note: The numbers do not add up to 100% because a comment could be assigned up to three comment themes.

Figure 5: Percent of Comments Containing Theme - Question 85 (Thinking about all the needs of your district, please discuss the two or three biggest advantages to implementing computerized testing.)



The three most prevalent themes present in the comments for Question 85 (Thinking about all the needs of your district, please discuss the two or three biggest advantages to implementing computerized testing) were Feedback, Electronic Testing (eTesting) Process, and Test Security. Within the Feedback theme, comments appear to center on the speed of receiving scores and the advantages of using feedback for planning. Within the Electronic Testing (eTesting) Process theme, comments center on the advantages of computerized testing versus paper and pencil testing; including streamlined processes, easier material collection, and less paperwork. Within the Test Security theme, comments focus on increasing or maintaining secure tests.

Illustrative comments – Question 85

Many of the comments were thoughtful and instructive. Comments often enhance the information value of the quantitative survey results. Comments enable respondents to provide input that is not specifically covered in the other (closed-ended) survey questions and to provide their own district perspective. Following each theme listed below are sample comments (from all the comments submitted) that illustrate the themes. Note: The comments are presented as typed by the respondents.

FEEDBACK

The ability to get the results in a short time. Less chance of error when checking the answers. Save the cost of print and shipping the tests.

The biggest advantage would be the quickness of computerized testing and the fast turnaround for results. Computer testing would also eliminate this mountain of paper we all have to deal with.

Quick turn around of test results

(1) Reporting to schools would be much faster. (2) Students seem to stay more focused when testing on the computer.

1. No materials to handle 2. Faster results

ELECTRONIC TESTING (ETESTING) PROCESS

Quicker feedback, a streamline testing process and lower human errors

Less paperwork, eliminates the signing in and out of test materials, reduce delivery cost, no packaging of materials, less manpower, and more directly dictates accountability.

If we could maneuver through the technology issues, then getting scores back more promptly would be wonderful. The sorting, packaging and shipping would be eliminated.

1. reduced amount of paper, 2. efficient scoring, 3. electronic data available for aggregating, 4. turn around time for scoring much faster than paper testing, 5. in general, the availability of data is much greater with computerized testing.

The biggest advantage would be the quickness of computerized testing and the fast turnaround for results. Computer testing would also eliminate this mountain of paper we all have to deal with.

TEST SECURITY

1. Quick turnaround of student results. 2. Reduction of test security issues involving testing materials. 3. Reduction of paperwork.

Less paperwork and security coding. Assess students at individual levels. Immediate feedback/scores. Greater accessibility for students. Better test security

Data stored out of the district may be more secure. Test materials are kept to a minimum.

1. Reduction in paperwork for statewide testing. 2. Increased test security.

Security of testing, not having to manually handle and secure all of the tests and their results.

TIME

Time Savings issues with Testing coordinators and accurate reporting.

1. Time spent on testing/shipping, collecting test materials. 2. Return of results for planning and implementing for the new year. 3. Security would not be necessary.

EFFECT ON STUDENTS

1. The speed with which we would get results. 2. Students are accustomed to computer testing already since we test with MAP three times a year.

Limit the amount of paper being used. Get students in the habit of using computers. Reduce human error.

PERSONNEL

Huge reduction in overhead of teachers required to administer and proctor tests as well as a reduction in the amount of time required to process the tests. Less overhead on paper production as well as any physical media required to present paper based tests. Test results can be packaged and presented in near real time via web or email, as opposed to teachers handing back test sheets a day or so later.

Quick turnaround of results Less work on school test coordinators

COST

Less paperwork, eliminates the signing in and out of test materials, reduce delivery cost, no packaging of materials, less manpower, and more directly dictates accountability.

I do not receive a budget and only get fund from the SCDE each year that has to do me for all software and hardware upgrades. If we do not receive any funds, then there is nothing upgraded. Also, I have to support all software and hardware related issues and I have to maintain SASI, MAPS, etc. We need support staff that would be available during testing times.

Current Testing Programs

Our students are familiar with computerized testing (MAP testing).

Previous experiences MAP Testing three times a year for last five years

Technology

Hopefully standardization of hardware and testing software as suggested by SDE; timely access to student data, a state supported testing solution

1) Computerized testing provides fast turn-around of grading the tests. 2) Schools will have additional computer labs for daily use. 3) Less work for staff in keeping up and shipping of test booklets.

Special Needs

More useful data due to rapid turn-around potential Environmental benefits will save a lot of trees Reduction in number of coding issues that result in unmatched data Comfort zone for children is using technology More authentic testing potential to use technological tools that are used on a daily basis by special needs children

MISCELLANEOUS

none

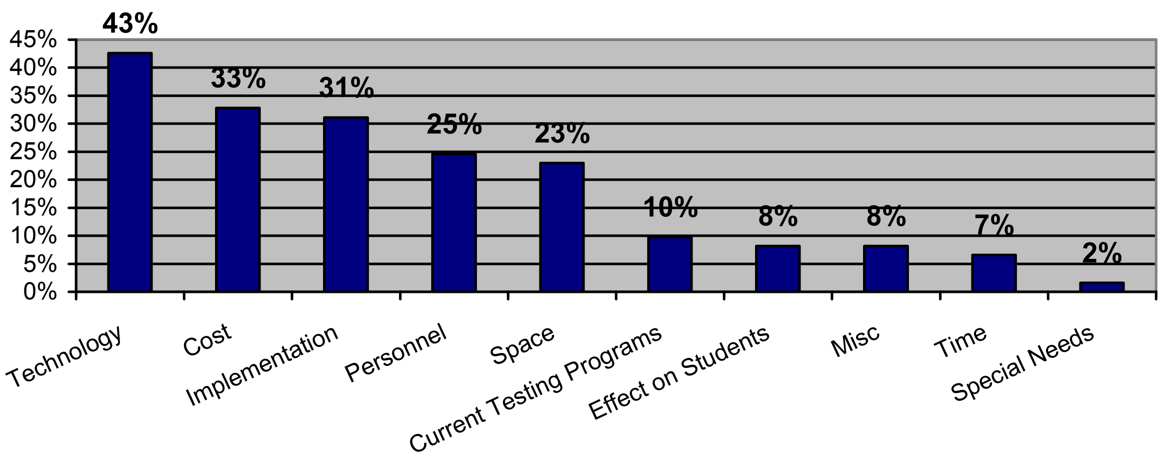
SPACE

Quicker test score results so adjustments can be made sooner. Additional labs could be utilized for instruction when not needed for testing.

Question 86

There were 61 comments provided in response to Question 86: Do you have any other comments about how your district could best transition to computerized testing? Figure Six displays the percentage of comments containing each comment theme in descending order from most to least frequent. Note: The numbers do not add up to 100% because a comment could be assigned up to three comment themes.

Figure 6: Percent of Comments Containing Theme - Question 86 (Do you have any other comments about how your district could best transition to computerized testing?)



The three most prevalent themes present in the comments for Question 86 (Do you have any other comments about how your district could best transition to computerized testing?) were Technology, Cost, and Implementation/Transition. Within the Technology theme, concerns appear to center on the degree of technological readiness for testing, including issues of whether schools needed to upgrade their technology in order to be prepared for testing; whether there was an adequate number of computers for testing and whether there would be sufficient bandwidth for testing. Within the Cost theme, concerns focus on funding for computer hardware, software, and personnel. Within the Implementation/Transition theme, concerns center on issues of scheduling computerized testing versus paper and pencil testing and pilot testing.

Illustrative comments – Question 86

Many of the comments were thoughtful and instructive. Comments often enhance the information value of the quantitative survey results. Comments enable respondents to provide input that is not specifically covered in the other (closed-ended) survey questions and to provide their own district perspective. Following each theme listed below are sample comments (from all the comments submitted) that illustrate the themes. Note: The comments are presented as typed by the respondents.

TECHNOLOGY

Wireless mobile labs will be the best way to move for our district. Could we implement this in one test or one grade at a time - with a plan?

1. Coordination of resources for testing at each school. 2. Establishing standardized hardware at each school for computers that are used for testing. 3. Standardizing of software on all computers used for testing.

Increase technology in rural communities.

Funds made available to build space, provide computer and personnel to support this effort.

Knowing the exact number of computers to have in place for testing to allow the student population test in a timely manner. It is critical to know this for planning.

COST

The state needs to provide proper funding so as to have up to date computer labs that could handle the testing.

We would need more technology funding to get our infrastructure and hardware up to date to accommodate online testing. Also, funding for additional technicians to cover the increased workload.

Funding needs to be significant and ongoing. Funding needs to include technology infrastructure, hardware, and support.

Equipment funding and physical space requirements are our biggest barriers to overcome to support the transition.

If state provide support through the funding of more software, hardware, staff and network and electrical infrastructure, computerized testing could become a reality.

Implementation/Transition

It would probably be easier for my district to transition to computerized testing if it were phased in by grades. It might be a good idea to start with the state mandated grades.

Wireless mobile labs will be the best way to move for our district. Could we implement this in one test or one grade at a time - with a plan?

1. A pilot project at various grade levels to determine significant testing issues and perform necessary adjustments regarding these issues.

I would hope that we work with a pilot program before full implementation takes place.

I would love to be a part of a pilot program

Personnel

Adequate training provided by an outside source would be critical.

Make training available.

SPACE

Equipment funding and physical space requirements are our biggest barriers to overcome to support the transition.

An incredible amount of resources is needed for hardware, additional personnel, and physical lab locations. We currently do not even have one technician per school, so additional personnel would have to be hired. Technology personnel would be needed at each site to administer the tests, train others, maintain test security, and ensure proper operations of both hardware and software. We would need to expose our students to computerized testing more. We can best transition being provided adequate funds to ensure successful test administrations.

CURRENT TESTING PROGRAMS

We are trying to transition now. We currently use NWEA MAPs and do some of the EOCEP online. We do not have enough un-encumbered labs at the high school level to test the number of students we would need to test. If we had a surge of new computers to solve the access problem, we would need help with the infrastructure to support it, we would need support monies to pay for technology support and our biggest obstacle for new labs would be space. Significant resources would be required to provide a space to house the new labs.

CURRENT TESTING PROGRAMS (CONTINUED)

Currently we administer MAP, a computer adaptive test, and our students are used to taking a test on a computer. However, the testing window and the capabilities of the system should be considered. The current system is very user friendly to our school staff. On another note, the current computer labs that are used for testing are also used for classroom instruction; we need to have more computers in the classroom. Ideally, we would like to have a computer for every student.

EFFECT ON STUDENTS

The state will have to change the way it approaches testing. There simply isn't physical space to test 100's of students at the same as we do with paper tests. The computers used for testing are an important part of our instructional program now and using them for testing will detract from all grade levels and students, not just those taking the tests.

This would be impossible for us to do because our district strictly prohibits inmate/student accessibility to the internet.

MISCELLANEOUS

Not at this time

N/A

TIME

Slowly

Additional resources would be necessary to transition to computerized testing. This includes and is not limited to funding for training, equipment, monitoring. Adequate preparation time is extremely important to a successful transition to computerized testing.

SPECIAL NEEDS

Consideration needs to be given to the needs of students with IEPs. How will their testing modifications and accommodations be met if a computer based format is used. Also, what are the implications for SC ALT?