Computer Based Testing in Utah: A Summary of Key Issues and Stakeholder Input

Marianne Perie & Scott Marion, Center for Assessment
February 29, 2008

Executive Summary

Utah is moving rapidly towards full implementation of computer based testing. The Utah State Office of Education convened a "summit" on computerized assessment in November 2007 to more fully explore the issues and options and to solicit input from Utah education stakeholders. This "white paper" is intended to inform Utah's long-term vision for computerized testing by providing both expert advice and district opinion. This vision articulates such issues as delivery methods, item formats, scoring approaches, reporting, integration with existing management systems, logistics, and practicalities.

The summit focused on discussing issues relative to computer-based testing (CBT) and computer adaptive testing (CAT). A CBT is a test administered by computer where all students receive the same items. A CAT is a test, administered by computer that dynamically adjusts itself to the achievement level of each examinee as the test is being administered. Advantages and disadvantages of each type were discussed. Other major discussion topics examined included:

- The types of items that can be developed for computerized tests that go beyond options for paper-and-pencil tests, such as interactive items or items with an audio or video component;
- Issues of access and accommodations for students with special needs;
- Infrastructure issues, such as:
  - connectivity and the physical location of the computers;
  - wired versus wireless networks,
  - administering the test on local software versus over the internet, and
  - laptops versus desktops;
- security concerns with computerized assessment and how those related to the timing, location, and type of test administered;
- reporting results in useful formats; and
- providing sufficient professional development to help teachers both prepare for the test and use the results appropriately.

The result of the summit and a subsequent meeting with a subgroup of participants was a list of decisions that need to be made about adopting a computerized assessment program. Not the least of these issues is to determine whether computerized assessment should be included only in the summative, accountability tests, or also adopted for interim and formative assessment.

Introduction

Utah is moving rapidly towards full implementation of computer based testing[1]. This is a major transition year with districts significantly increasing the numbers of students completing their criterion-referenced tests (CRTs) via computer. Further, the Utah State Office of Education (USOE) will need to issue a request for proposals (RFP) to hire a contractor as USOE shifts the entire testing system to a computer platform. The state's technical and policy advisory committees recommended that USOE convene a "summit" of Utah's key education stakeholders to articulate a long-term vision of computerized testing in Utah. The advisory committees made this recommendation to ensure that USOE understands this vision as it proceeds with short term implementation necessities. Hence USOE convened a summit on November 7–8, 2007.

This "white paper" is intended to inform Utah's long-term vision for computerized testing by providing both expert advice and district opinion. This vision articulates such issues as delivery methods, item formats, scoring approaches, reporting, integration with existing management systems, logistics, and practicalities. We hope this document is useful in the ongoing discussion regarding the future of testing in Utah and to inform the request for proposals. Before moving into the discussion of the summit, we present a brief history of computer based testing in Utah.

A Brief History of CBT in Utah

Utah's criterion-referenced testing (CRT) system includes 29 unique tests in mathematics, English language arts and science as part of Utah's Performance Assessment System for Students (UPASS). This does not include other accountability level assessments administered such as the direct writing assessment (DWA), the exit exam (UBSCT), Utah's alternative assessment for students with significant cognitive disabilities (UAA), the norm-referenced testing system, and the Utah Alternate Language Proficiency Assessment (UALPA) for the state's English language learners. All tests rely exclusively on multiple choice items, except the DWA, the writing component of UBSCT, and the UALPA, in part because of the large number of assessments to develop, administer, and score, most of which is done in house. When UPASS was first enacted, the CRT system was intended to be part of an overall system of focusing and improving learning in the core areas of English Language Arts, mathematics, and science. With this in mind, the assessments were developed to maximize reporting capability while being cognizant of administration constraints. This resulted in the ability to report at the standard and objective level.

In 2002, Utah began considering offering CRTs via a computerized testing system. The primary purpose at the time was to decrease the time to return results. There was also

---

[1] "Computer-based testing" is the common term, but because this phrase was used to mean a specific type of testing in the summit, we will be using the term "computerized testing" when referring to the concept in a more general manner.

a desire to explore the use of innovative item types in CRTs in order to improve the validity of the assessment, allow for the integration of multiple item types, and to increase student engagement. The first operational computerized tests in Utah were administered in 2004. The first couple of years of computerized testing in Utah saw relatively few students tested in this format (no more than 10%) By 2006, a single test vendor, under contract directly to the state, was being used although still relatively few students were being tested in this format. In 2007 the emphasis from policy makers on computer based testing increased dramatically. The objective was to move faster to return results and explore the use of innovative items on the CRTs and to utilize the system for an increase in formative assessments. The 2008 CRTs will see approximately 50% of the 1,200,000 test administrations occurring via computerized testing with the intention of moving as close to 100% as possible in the next couple of years.

The Utah Test Item Pool Service (UTIPS), an important aspect of computerized testing in Utah, has been in use since 2000. This offers assessment items, delivered via the USOE website, aligned to Utah content standards for teachers, schools, and districts to use for formative and interim assessment purposes. This system offers a variety of assessment item types including constructed response items and performance tasks. This computer based system provides classroom teachers with items and tasks to use as part of classroom formative assessments. The system offers both intact assessments and items with tools to help create interim/benchmark assessments. USOE has documented the high levels of satisfaction and growing use of the UTIPS system by Utah educators.

Description of Summit Attendees and Process

Leaders from all Utah school districts, including charters, were invited to send their assessment and technology directors and others they thought would contribute to the outcomes of the summit. Several members of Utah's state board of education and Utah's legislature were invited and attended the summit. Many members of USOE's assessment, accountability, and computer services units also attended the summit. Including the three outside experts and the five staff from the Center for Assessment, there were over 130 participants at the two-day summit.

The summit was structured to allow time for presentations from national experts in computerized testing followed by discussion from attendees. There were separate breakout sessions for each of the four key areas of computerized testing: item types, technical issues, infrastructure, and reporting and integration. The discussion in the breakout groups was intended to elicit the opinions of Utah education stakeholders on these key issues (the summit agenda is found in Appendix A). There is also a fifth key issue — access and accommodations. A short presentation was given on this topic, but no separate breakout session was held, so no opinions from Utah educators are expressed on that topic here.

The summit organizers and the Center for Assessment staff felt that it was important to include experts in computer-based testing to help provide a foundation for the key

issues. Jim Olsen, Ph.D., Senior Associate of Alpine Testing Solutions is internationally known for his work in computer-based and computer-adaptive testing. He was a coauthor of a seminal chapter in the 1989 edition of Educational Measurement (3rd edition) about computer-based testing. Neal Kingston, Ph.D., a professor at the University of Kansas, has both the deep understanding of an academic researcher and the practical experience as a former testing contractor (Dr. Kingston was the Vice President of Measured Progress and Director of Research for CTB/McGraw-Hill) that he brings to bear when talking about computer-based testing. In addition to his recent research about comparability of scores between computer- and paper-based administrations, Neal has been an instrumental partner in the effort by several states to implement computer-based testing. Cynthia Parshall, Ph.D. is an internationally known consultant on computer-based testing and has become known for her ability to translate complex ideas into useful practical approaches for implementing computer-based testing. Her framework for item types, for example, has been used by many test developers and others when designing computer-based testing programs.

Organization of the White Paper

In general, this white paper follows the organization of the summit. The first major part of this paper presents an overview of the five key issues with an intention of outlining some of the key considerations of these major facets of computerized testing for Utah policy makers. We also discuss levels of assessment—formative, interim, and summative—to reflect the amount of time spent discussing this topic at the summit. The second major part of the paper summarizes the attendees' opinions and interests around the key issues. Finally, the paper concludes with a summary of the long-term vision and recommendations for next steps that USOE may consider.

PART I: SUMMARY OF EXPERT PRESENTATIONS

Many of the discussions about computerized testing returned to the issue of the usage of the test. The participants had different viewpoints depending on whether the computerized testing was to be used for accountability testing, interim testing, or formative testing. Therefore, this paper begins with a discussion of these three assessment types, which will be referenced throughout the remainder of the paper.

Different Uses for Assessment

The Center for Assessment has spent considerable time over the past two years distinguishing among assessments used for formative, summative, and interim purposes. In their recent paper, Perie, Marion, and Gong (2007) proposed the following definitions to distinguish among the three main assessment types:

> Formative: *Formative assessment is a process used by teachers and students during instruction that provides feedback to adjust ongoing teaching and learning to improve students' achievement of intended instructional outcomes.*

Interim: *Assessments administered during instruction to evaluate students' knowledge and skills relative to a specific set of academic goals in order to inform policymaker or educator decisions at the classroom, school, or district level. The specific interim assessment designs are driven by the purposes and intended uses, but the results of any interim assessment must be reported in a manner allowing aggregation across students, occasions, or concepts.*

Summative: *Summative assessments are given one time at the end of the semester or school year to evaluate students' performance against a defined set of content standards. These assessments are typically given statewide (but can be national or district) and are usually used as part of an accountability program or to otherwise inform policy. They could also be teacher-administered end-of-unit or end-of-semester tests that are used solely for grading purposes. They are the least flexible of the assessments.*

Given these definitions, which focus on the use of the assessment, UTIPS could be considered either interim or formative. Because UTIPS contains a collection of items in an item bank, a teacher could create an assessment and ask her entire class to take it and use it to evaluate their progress on the objectives to date. That would be an example of an interim assessment. However, the teacher could select different items for different students to take when they were ready to determine the depth of their understanding or to create a task to challenge them within the unit. These would be examples of formative use. At the other end, the current CRT assessments are summative assessments used for accountability purposes. It was important for participants to be clear on which purpose they were considering for computerized testing. In the opinions expressed in Part II of this paper, it is clear that the participants had different recommendations for accountability tests than for either formative or interim tests.

Type of Computer Based Tests

Before beginning with a serious discussion of the issues surrounding computerized testing, it is important to come to a common understanding of the definitions of each type of test. Although "computer based testing" or CBT is typically used as an umbrella term to describe all types of computerized testing, that phrase is also used as a specific type of computerized testing. In addition, our experts discussed computer adaptive testing (CAT) extensively and also touched upon continuous measurement and intelligent measurement.

Jim Olsen gave definitions for the four different types of computerized testing as follows:

*A **Computer-Based Test (CBT)** is a test administered by computer. All students receive the same items. The CBT can be delivered via network or online.*

*A **Computerized Adaptive Test (CAT)** is a test, administered by computer that dynamically adjusts itself to the trait level of each examinee as the test is being administered. The CAT can be delivered via network or online.*

***Continuous Measurement**: uses calibrated items and tasks to continuously and unobtrusively estimate and profile dynamic changes in examinee proficiency.*

***Intelligent Measurement**: application of knowledge-based computing to any of the processes of educational measurement (test specification and item banking tools, intelligent scoring of complex constructed responses, automated profile interpretations, intelligent tutoring within job element or task, automated task generation). "Assessment Engineering"*

The last two types were not discussed beyond these definitions. However, there was extensive discussion comparing and contrasting CBT and CAT.

Computer Based Tests (CBT)

The experts at the summit discussed many advantages to using CBT. Some of the advantages were dependent on the purpose of the assessment. CBT could reduce the teacher load on grading selected-response items on interim assessments, for example, and produce score reports quickly. It would also allow teachers to select the items they wish to use and produce their own test relatively easily. Regardless of the purpose of the assessment, using a CBT could reduce the turnaround time for producing results, particularly when a pre-equating[2] design is used. The use of CBT also increases the

---

[2] Equating is the process of mathematically relating the scores from one test form (e.g., the 2008 4th grade math test) to another test form designed to measure the same knowledge and skills (e.g., the 2007 4th grade math test). Most designs rely on a "post-equating" approach where the mathematical relationship between the two test forms is not established until after both administrations are completed and is based on both sets of students completing a set of linking or "anchor" items. Pre-equating, on the other hand, is based on having students simultaneously respond to all of the items that could be used on any future test, with the items randomly distributed across students in a pilot test. Then those items and the associated statistical characteristics are stored in a pool and used to create two or more equivalent test forms in the future. There are many testing programs that rely on pre-equating, but the quality of the pilot test is critical to successful pre-equating. Post-equating is often used to validate pre-equating results which may leave states with the challenging decision of whether or not to adjust scores or scales if the two approaches yield different results. Finally, while the long-term costs of using pre- and post-equating approaches are likely equivalent, the up-front costs of shifting to a pre-equating approach are considerable because of the need for shifting a large portion of the item development and pilot testing to the first year or so in the program instead of spreading it out over four or five years.

range of item types possible, although it would be important to ensure that all students in the state have access to the same types of items. The feasibility of using some types of items may be dependent on the quality of the hardware available as discussed in the section on item type. For example, CBT provides the opportunity for administering unique performance items, such as simulations or responses to video and audio files.

CBT also can measure different aspects of student performance such as response time and answer changes. These types of information can be useful to teachers as additional diagnostic information. CBT could eliminate the construct-irrelevant variance[3] caused by students "bubbling in" the answer on the wrong line of the response document. Students answer the question as they see it, and most programs prompt the student to "be sure" of their final answer. Finally, one expert indicated that there is some evidence that students enjoy taking tests more on a computer, although there was some speculation that this effect may decrease over time as students become accustomed to computerized testing.

There were also some disadvantages to using CBT. The same items are given to every student, maintaining many of the same issues of cheating as paper-and-pencil tests and raising the issues of sufficient computer time. Likewise, the CBT has the same issue of paper-and-pencil tests of being unable to adjust to a student's ability level, possibly resulting in a test being "too easy" for some students, while causing other students to be frustrated time and time again as they are presented with test questions beyond their ability level. Moreover, to fully maximize the benefits of quicker turnaround time for producing scale score results, a pre-equating approach will need to be implemented, which carries certain challenges, risks and upfront costs.

One benefit of CBT that may not be as attainable with CAT is the use of the assessment for diagnostic purposes. If the items are designed in a manner that allows incorrect responses to convey important diagnostic information, then having every student take the same item can provide valuable information to a teacher, particularly when the CBT is used as an interim assessment.

Computer Adaptive Tests (CAT)

Computer adaptive tests (CAT) represent a new way of thinking about assessments, not simply a new platform for administering them. They therefore have additional complexities. Because CATs are relatively new to Utah, much time was spent discussing

---

[3] Messick (1989) discussed construct-irrelevant variance as a major threat to validity. The "construct" is what test specialists refer to as the target of the test. It is the theoretical trait (e.g., reading achievement) that we are trying to measure with the test. Construct irrelevance occurs when factors other than what we are intending to measure affects the test score for a student. A common example is when a "science" test has such a heavy reading load that it really becomes more of a reading test.

them in depth. This section will capture the expert presentations by first describing the advantages of CAT and cautions for using them, followed by presenting some technical considerations.

Advantages of CAT

One advantage of using CAT is that it can maximize information while reducing testing time, thus increasing efficiency. According to proponents, CATs can reduce the test length by 50% or more compared to conventional tests. CAT capitalizes on advances in psychometric theory. For instance, the integration of item response theory in the development of CAT allows the assessment designers to control directly the measurement precision. Unlike conventional assessments, a properly designed CAT can measure or classify all examinees with a certain degree of precision. Moreover, the CAT can provide good information at all (or at least most) points on the full score continuum. Thus, the use of CAT allows for an efficient determination around a specific decision point (e.g., Proficient/Not Proficient).

The CAT also allows for greater test security. For instance, items are more secure as different students see different items, reducing the potential for cheating. This factor also allows more flexibility in administering the tests, as students can more easily take the test at different times without risk of cheating.

Disadvantages of CAT

Perhaps the most challenging aspect of working with CAT is explaining it to parents and teachers. Each student takes different items and even a different number of items, which can be difficult to explain to teachers and parents. Because the student is tested until a specified level of precision is obtained, raw score statistics such as number right and percent correct lose their meaning. Two students could score the same percent correct or have the same number correct but have very different scale scores. This concept can be difficult to convey to non-measurement experts.

Secondly, ensuring alignment to the content standards can be much more challenging with CAT. Unless CAT is designed to measure each content standard to a specified level of proficiency, the focus is on the total score in the subject. Thus, reporting standards-based interpretations and producing standards-based diagnostic information can be significantly more challenging with CAT.

Technical Aspects of CAT

CAT requires the use of item response theory (IRT) to relate item and task performance to ability estimates and proficiency classifications. There was discussion about the technical issues associated with using a CAT centering on the five components of CAT:

1. **A pre-calibrated item bank**, which requires developing and adding large numbers of items often to maintain a bank of items with a range of item parameters to approximate the desired target bank information function

2. **A starting rule for selecting the first item**, such as the desired difficulty level of the item

3. **A procedure for scoring item responses and estimating trait level,** typically by estimating the ability level of an examinee using IRT, selecting an item (or item set) around that ability level, and then revising the ability level and standard error based on the response to that item (or item set).

4. **A method of selecting the next item**, for instance selecting the item or item set that maximizes the test information or selecting the item or item set that minimizes the standard error.

5. **A rule for ending the test**, such as when the standard error of measurement around the student's ability level falls to pre-determined maximum allowable level or when the ability plus/minus two standard errors of measurement all fall within the same performance level.

Dr. Olsen also presented different administration models. *Item-by-item CAT* adapts or tailors the difficulty of the test to each examinee, item by item. Items are selected to maximize test information and minimize the standard error of measurement of the examinee's ability estimate (θ[4]). *Classification CAT* is based on either item or testlet selection. The final score is a classification score indicating whether the examinee score and confidence band are above or below the designed performance standard. Testing proceeds until the θ estimate and confidence band are completely below (fail status) or completely above (pass status) the designated performance standard.

A brief explanation of testlets is needed. The testlet is a structured sequence or cluster of items that can be used as the unit for CAT item selection and scoring. They can be organized by content standard and the test structured to ensure that students answer items covering an appropriate range of content. That is, there may be five testlets within a panel (the entire scored test) each covering one of the five math content strands. A student will be administered each of the five testlets, and the computer will calculate an ability estimate for each strand as well as an overall math ability. Using a testlet structure helps test designers ensure an adequate range or breadth. They can also be used to ensure that students who miss an "easier" item still have the opportunity to answer more challenging items until the ability estimate is ascertained within a reasonable level of certainty.

The final technical issue that was discussed related to equating designs. Equating is the process of putting the scores from the current year's administration on the test scale

---

[4] In Item Response Theory (IRT), the examinee performance that we are trying to estimate has traditionally been referred to as "ability" (based on the aptitude testing tradition) and is denoted as the Greek letter theta (θ).

from the previous year so that results across the two years can be validly compared. Although both CBT and CAT need to be designed with equating in mind—meaning all forms should measure the same construct with the same reliability—there are some issues specific to CAT. CAT requires pre-equating, so that all items in the item bank are already equated and placed on a single scale. This is a necessary component of CAT and allows for faster turnaround time for producing scores for all computerized testing. Although pre-equating will also reduce turnaround time for CBT, it is not a necessary condition for using them. As with all testing programs designed to measure change, ensuring the implementation of valid equating designs is critical.

Most state summative testing programs rely on post equating designs where the item statistics used to link the current test to the previous year's test scale are derived from the actual operational test administration. Pre-equating designs, on the other hand, generally rely on the item statistics derived from the field test so that the test specialists do not have to wait until after the test has been administered to produce equated scale scores. Therefore, the quality of the field test is critically important. Post-equating is often conducted to check the accuracy of the pre-equating. When differences between the pre and post equated scores are uncovered, the users (the state in this case) is left with the choice of re-issuing scores or more commonly adjusting the trend line before the next year's test is administered. Of course, this means that less than accurate scores are used for the given year.

Developing an Item Bank

Using computerized testing requires developing an item bank which can be used to create test "forms." Cynthia Parshall presented ideas of ways to extend item types beyond simple multiple-choice or short answer items that are typically seen on paper-and-pencil tests. Her framework includes seven components: assessment structure, complexity, fidelity, interactivity, response action, media inclusion, and scoring algorithms. This section will provide an overview of this framework and also touch on issues of item development for computerized testing. In general, each dimension provides an approach for assessing student understanding that may improve the measurement of the construct but also may increase the cost.

Assessment Structure

Assessment structure translates to item format. With computerized testing, an assessment developer has many more options in terms of the type of item format. Not only can students respond to traditional multiple-choice and open-ended items,[5] but these items can take on new formats. For example, a multiple-choice item might be a diagram, and the task might be to click on the appropriate symbol or position given a choice of four symbols or positions in the diagram. The task in a science test may be to

[5] An "item" is a test question or a specific task to be performed as part of a test.

"click on the area that best represents the area in the earth where magna may be found." A constructed-response item might include a figural response asking a student to drag and drop the appropriate mathematical expression to complete a formula. Or the item could be embedded in situated tasks. A situated task is an integrated, realistic situation or scenario in which examinees are asked to solve a problem or task, typically through a series of actions or steps, like an architectural design. Or, tasks may be given in simulated environments presented in realistic settings that replicate the experience of interacting in a real environment. A flight simulator is an example of a simulated environment. More innovate assessment structures can improve the measurement of a construct. However, they require custom programming which increases the cost, and they rely on the examinees having greater computer skills. Furthermore, much less is known about the psychometric functioning of innovative item types.

Complexity

After item type, the assessment developer should consider the complexity of the computerized testing

Complexity refers to the number and variety of screen elements that an examinee needs to interpret. This could represent construct-irrelevant variance and needs to be monitored closely. For example, multiple headers, text boxes, and portion of graphics that are clickable can greatly enhance the difficulty of a task. Also included under complexity is the number of onscreen elements that an examinee can use, such as response options, graphical tools, navigation buttons, and audio/visual players. The complexity can range from a relative low complex task, such as one with a traditional four-option response, to a moderate task of selecting an element from a workshop, to a highly complex task such as completing a form. More complex item types may improve the measurement of the construct but may also increase construct irrelevant variance. Poor user interface can increase the measurement error. More highly complex tasks also require custom programming, which in turn will increase the cost.

Fidelity

Fidelity is the degree to which the assessment provides a realistic and accurate reproduction of the actual actions, tasks, and environments that are part of the construct being measured. The closer the assessment approximates the actual construct being measured, the higher the fidelity. An example of a low fidelity item is one that includes a text-based scenario and asks a multiple-choice question about what the student read. A high-fidelity item would be one that asks the examinee to reply to an email and actually send it. Again, high fidelity can improve the measurement of the construct, but it can also require special programming and be costly.

Interactivity

Interactivity is the degree to which an assessment responds to examinee actions. For example, a low degree of interactivity could include "clicking" an on-screen figure, which displays a single, contextualized result, such as an English 'editing' task. An

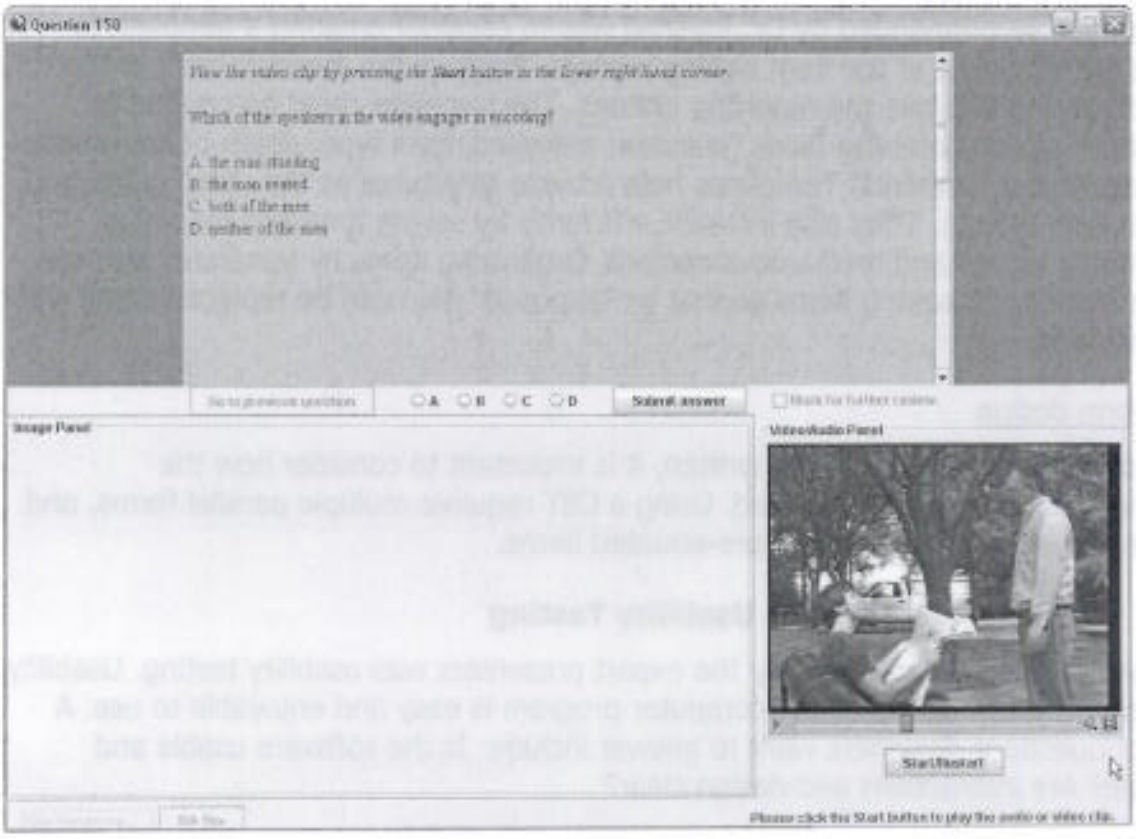
example of a task with moderate interactivity would be one in which there were two or three iterations of examinee/computer interactions, such as asking a student to perform a sequence of tasks to graph the results of a science experiment. For high interactivity, consider the potential for a series of examinee/computer iterations, such as with an interior design task. Increasing the interactivity may improve the measurement of the construct, but again will require custom programming and may be more expensive to develop. In addition, increased interactivity requires great computer skills of the examinee and additional psychometric analyses by the assessment developer.

Response Action

Response action includes the actions the examinee is asked to take and the input devices an examinee is asked to use. Examples of input devices include a keyboard, mouse, touch screen and microphone. Allowing the use of more innovative response actions may improve the accessibility of the task to all students and may improve the measurement of the construct. However for highly unusual response actions, additional psychometric analyses may be needed. Occasionally, custom programming will be needed to enable the CBT to utilize the input device/response action. And, again, expenses might increase, due to the need to purchase alternative input devices or to develop tutorials

Media Inclusion

Media inclusion refers to the addition of non-text material in the item stem and/or response options. Some examples of media inclusion would be graphics, animation, audio, and video. The screenshot below shows an example of an item that incorporates a video clip into the test question. Again, using these various forms of media may improvement the measurement of the construct. However, custom programming may be needed for some media types, and more innovative uses of media may require additional psychometric analyses, all of which can increase the cost. In addition, some media types will have implications for bandwidth, test delivery hardware, etc.



Scoring algorithms

The scoring approach could be dichotomous (e.g., selected response), polytomous (e.g., open-ended), or involve complex modeling. A benefit of using a scoring algorithm is that more elaborate scoring algorithms may increase the amount of information obtained from the examinee's response. However, the development and validation of the scoring model may require substantial psychometric effort. Additional programming may be necessary to enable the collection of the examinee's response, and expenses might increase, due to psychometric and programming needs.

Item Writing Guidelines

Given the various opportunities and complexities of developing items for computerized testing, some different considerations for writing items might be needed. As with traditional assessment, it is important to specify the appropriate content and grammar of the items. However, the format and layout of the items has more possibilities when using computerized testing. But more possibilities also lead to greater complexities. For instance, when using an audio file to model a conversation, it will be important to consider the number of voices a student can follow. Test designers will need to improve bias and sensitivity reviews to consider the accessibility of the item. It will be important for reviewers to consider access and accommodations, particularly in terms of the complexity of the item, where complexity refers to the number of on-screen elements the student must use to answer the question.

One approach recommended by Dr. Parshall was to consider creating a template to increase the efficiency of the item writing process, improve the measurement of the construct, and to improve the reporting options. The template could be created to include both typical database fields (standard assessed, item type, depth of knowledge) and screen layout elements. Templates help provide structures as they both guide and constrain item writing. They also increase efficiency by saving time and money in programming efforts and media development. Organizing items by templates also can increase security by pairing items so that an "exposed" item can be replaced easily with a close variant.

Form design

Finally, once the items have been written, it is important to consider how the assessment forms will be assembled. Using a CBT requires multiple parallel forms, and a CAT requires an item bank with pre-equated items.

Usability Testing

One issue stressed as important by the expert presenters was usability testing. Usability describes the extent to which the computer program is easy and enjoyable to use. A couple of questions designers want to answer include: Is the software usable and accessible? Are instructions and design clear?

We do not want to introduce any construct-irrelevant difficulty into a test based solely on the student interface with the computer program, as it would lessen the validity of the results. We also need to be concerned about accessibility across alternate platforms and hardware. There needs to be consistency across computers, classrooms, schools, and districts. One recommendation from the presenters when shopping for systems was to require anyone providing a system to prove that it can do what they claim it can do. That is, ask for the data showing the usability of the system.

Methods for Doing Usability Testing

There are several ways to test the usability of the system. One is to apply think-aloud protocols with the examinees. That is, have the examinees say out loud what they are thinking as they interact with the computer during testing. Then, determine how closely their thinking is aligned with the construct being tested versus trying to figure out the computer program. This would also include observing how long it takes for students to move from reading the instructions to interacting with the item. Structured interviews with the examinees after they have taken the computerized test can also provide valuable information on the ease of use of the system. Asking an expert to review the system for usability or testing it yourself with a usability checklist can also provide some valuable information.

Access & Accommodations

Scott Marion from the Center for Assessment provided a summary of some of the most recent research on access and accommodations issues as well as an expert analysis from the National Center for Educational Outcomes (NCEO) about the interaction of computerized testing and access for students with disabilities. In general, technology can improve access and achievement for students with disabilities because designers can program many accommodations directly into the testing program. According to Greenwood and Rieth (1994)[6], the primary strength of computer-based testing is its "potential for removing traditional barriers to the inclusion of persons with disabilities in the assessment process through adaptations and accommodations as well as through new forms" (p. 110). Therefore, in many circumstances, students may need fewer accommodations with computerized testing as the use of the computer is an appropriate accommodation or the computer may easily support an adaptive device. The biggest challenge for providing accommodations via computerized testing appears—for the time being—to be for students requiring Braille test forms. While the technology for such options is rapidly improving, the need for paper Braille forms will likely remain at least through the next contract period. Designers may also need to consider how a student with a hearing disability would access an audio file, if that is part of the assessment. Appendix B presents several tables from excerpted from NCEO researchers Thompson, Thurlow, Quenemoen, and Lehr (2002) outlining the degree to which a broad range of accommodations are a built in feature of universally designed computerized testing (which would be available for any student), the need for the accommodation is at least not affected by the use of computerized testing, or finally whether a new or different accommodation may be needed for computerized testing. It is important to keep in mind that this NCEO summary was written six years ago and the range of available built in accommodations has likely improved considerably.

A recent summary of the research (Zenisky & Sireci, 2007)[7] indicates that, in general, it appears that bundling accommodations with computerized testing leads to either improved outcomes for students with disabilities (and to some extent ELL) or no effect. In this case, "no effect" is a good finding. On the other hand, some researchers have raised legitimate concerns that several features may not be universally accessible because advancements in assistive technology are usually several steps behind new Internet components and tools. As computer-based testing becomes a reality across states and districts, it is important to ensure that the new technology either improves accessibility or is compatible with existing assistive computer technology.

---

[6] Greenwood, C.R., & Rieth, H.J. (1994). Current dimensions of technology-based assessment in special education. *Exceptional Children, 61*(2), 105-113.

[7] Excerpted from Zenisky, A. & Sireci, S. (2007). A summary of the research on the effects of test accommodations: 2005-2006. (Technical Report 47). Minneapolis, MN: National Center on Educational Outcomes. Retrieved 10/30/07, from: http://cehd.umn.edu/nceo/OnlinePubs/Tech47/TechReport47.pdf

Finally, to emphasize a theme that is found throughout this paper and a long-held view of those supporting increased access for students with disabilities and English language learners, accommodations provided via computerized testing during testing must be available for students during instruction well before assessment. If the first time a student sees a certain type of accommodation—especially if it is somewhat novel—is on the large-scale assessment, that accommodation will likely hinder instead of improve access. Further, technology cannot take the place of content mastery. No matter how well a test is designed, or what media are used for administration, students who have not had an opportunity to learn the material tested will perform poorly. Other challenges that must be overcome in order for computer-based testing to be effective include: issues of equity and skill in computer use, added challenges for some students, technological challenges, security of online data, lack of expertise in designing accessible Web pages, and prohibitive development cost.

Reporting

USOE will need to give considerable attention to designing the reports produced by computer-based assessment systems. When pre-equating is used, reports can be generated instantly with scale scores. This would have the added benefit of eliminating the need for releasing raw scores to the public, which are often easily misinterpreted. In these cases, performance level information can also be generated immediately. And while scale scores and summaries of items answered correctly and incorrectly may be available instantly, comparative information, such as classroom or district means or normative information on percentiles, will not be available until all students in the state have finished taking the test.

Additionally, computerized testing may lend itself to more diagnostic reporting. Information about the length of time it took to answer the item or about an incorrect response may provide a teacher with important information about student understanding. Depending on whether the computerized test is used for accountability or more formative purposes may also drive the design of the reports and the timing of the release.

Infrastructure

One session of the summit was devoted entirely to infrastructure issues. This topic covered issues such as the hosting of the assessment — whether it was housed locally or centrally via a website — connectivity, and technical requirements such as bandwidth and electrical power needs. Each of these issues will be summarized here.

Types of clients

Dr. Kingston introduced the terms, thin clients, thick clients and smart clients. "Thin client" means that the software is housed on the internet or at another location. The student accesses the test at a home terminal, but both the test and student responses are stored on a different server outside of the building. A thin client allows student data to be uploaded instantly and also allows upgrades to be downloaded automatically. It

reduces the burden on local technology support staff as they do not have to "touch" every computer when updating software. However, a thin client can be complicated to develop and requires a "web-based" technical support person with access to the program on the server to be available at all times. If the internet connection is lost, the student will be unable to take or complete the assessment. Issues of data security may also be a concern. Student data will need to be backed up locally, and blockers will need to be installed on the computer to prevent a student from finding an answer on the internet when accessing the internet to take the test.

Conversely, the term "thick client" means that the software is loaded directly onto the computer the student is using and that the student's responses are also stored on that computer. There is no dependency on another server outside of the school or on the availability of the internet. It is relatively easy to run, but a thick client requires a local technical support person to update software on each machine and backup data. It also may require schools to invest in a specific type of computer or operating system.

The third type of client is the "smart client" in which the software is housed both on the internet and locally. The smart client allows for both online and offline support. If the internet connection is lost, the student can continue working on the local system and the data will be saved locally. Then, when the connection is restored, the data can be uploaded easily. Likewise, it is relatively simple for a website manager to deploy and download changes. It does not require a local technical support person to install updates on each machine. However, it could restrict the flexibility of districts to change platforms. The smart client will need to be robust enough to handle changing vendors, software, or other conditions.

Connectivity

Another issue discussed was whether the computers should be connected to the internet wirelessly or hard wired. A wired connection is more secure and does not require encryption, but it can result in a lot of wires and depending on the age of the building, could lead to safety issues. In addition, it reduces the mobility of the computers. A wireless connection requires data to be encrypted, which can slow connection speeds, and creates wireless management issues, but it creates cleaner workspaces and more flexibility on where to place computers.

Along with the issues of connectivity, there are two main issues about the school building that can influence the potential for various types of internet. The bandwidth available and the electrical power needs should be considered, particularly for older schools. In addition, schools will need to consider both how computers are networked together within the school and how they are connected to the internet. A larger bandwidth will increase the connectivity speed. However, there needs to be a backup system in place to ensure that if a connection is lost, the information on the items that the examinee has already taken is not lost. In addition, sufficient electrical outlets are needed to support the use of computers in the classroom. Moreover, older buildings

may need to be rewired before they can support multiple computers, printers, and a server.

Practical and Logistical Issues

Several practical issues were discussed during the summit. Much time was spent early on discussing test windows and schedules. Dr. Kingston also presented information regarding the physical location of computers, either within the classroom, in a separate computer lab, or on mobile carts. Both of these issues are described in this section.

Test Windows and Schedules

Much was made of the fact that computerized testing could reduce the turnaround time for producing scores on the accountability tests, thus allowing for a later testing window. We must note that Utah turns around raw scores exceptionally quickly, almost always within four-five working days. However, the clock starts "ticking" for this four-five day timeline once all the tests are received by USOE's computer services department. Computer based testing will allow each student to receive essentially instant raw score reports. Instant or rapid scale scores can only be produced when the items are pre-equated.

The physical location of the computers is a major infrastructure issue and one in which the participants had significant interest. The experts discussed the issue of incorporating computers into instruction when they are located in the classroom.

Computer labs were presented as another possibility for using computers within a school. Labs allow many students to be on computers at the same time. However, teachers need to reserve time in the lab if they wish to bring their entire class. A lab setup might allow a teacher the freedom to send students to the lab when they are ready for the test while he/she continues to teach the remaining students in the classroom assuming that there is another adult in the lab who can monitor the students. This assumption of additional personnel may not be valid for all schools. Labs could also be used as a place to send students who are struggling with a particular unit for additional support, again assuming an adult is present to supervise the student in the lab. Schools will need to consider housing to determine if sufficient space exists for the labs. The layout will need to be considered as well to determine what type of organizational structure is preferable and whether or not it is necessary to build carrels within the lab to allow students to work independently and undisturbed. Student will need to have sufficient workspace for their keyboard, a mouse, and scratch paper or notes, if appropriate.

If there are many computers in each classroom—enough for testing—or with a lab environment the layout of the computers in the room is critical when shifting from instruction to large-scale assessment. In instructional situations, it is important for the teacher to easily be able to see many screens at once, but this also means that students can see many screens at once. Also, students need to be able to see both their

monitors and the teacher simultaneously. Dr. Kingston demonstrated how vertical monitors, which are very useful in instructional settings, can be both a distraction and a source of cheating. Dr. Kingston proposed several strategies to reduce cheating, such as using polarizing screens and organizing room on diagonal to allow the students to face the teacher without being directly behind another screen.

A third possibility for locating computers within the school is on mobile carts. Maintaining the computers on carts would allow students to use them either in a classroom or a lab. Carts generally work better in wireless situations where the computer does not need to be plugged into an Ethernet connection to access the internet. It is also easier to work with laptops on mobile carts than desktops because of the number of components and the weight of a desktop computer.

Security Issues

There are different security issues for computerized testing. For instance, issues of cheating have different implications. Item security may be similar for computerized tests as for paper-and-pencil tests, but there is the added concern of keeping secure student responses to tests. Issues such as back-up systems and encryption are unique to computerized testing. Each of these concerns is discussed on the following pages.

Cheating

Although cheating is an issue in paper-based assessments, as well as computerized testing, it is important to examine some of the concerns unique to computerized testing and consider how to reduce the risk of cheating.

The first issue to consider with cheating on computerized testing is the timing of the test. With CBT, students within a classroom will need to take the test simultaneously to reduce risk of cheating. Because the same items typically are given in the same order to all students, it is relatively easy for one student to inform another as to the test content. However, simultaneous testing may raise logistical concerns regarding the availability of computers as mentioned in the previous section. In addition, the layout of the computers is important to preclude the possibility of one student seeing another student's screen. Each classroom or lab would require careful monitoring.

The second issue with cheating is connectivity to the internet. One could foresee a situation where a student searches for the correct answer using an internet search engine while in the middle of taking a web-based assessment. Security provisions can be implemented to prevent students from visiting other websites while taking the assessment. For instance, blocks can be installed that allow students to access only the testing site during testing and disable all other sites.

Data Security

In addition to concerns about the security of the items, computerized testing requires attention to be paid to keeping student data secure. The data need to be secured both

from loss and from exposure. To ensure that student test data are not lost, they will need to be backed up and stored in multiple locations. Typically, student data are stored locally when the test is administered and then uploaded to a central site. Dr. Kingston recommended not deleting files from the local system until after (perhaps well after) notification that the central site has received data. This strategy may require additional local storage. Secondly, backup systems are needed to avoid losing test data in the event of a system failure during the test administration. If the test is administered via the internet and the connection is lost during the test, a backup system should be in place to prevent the loss of all student work done to that point.

The second consideration is keep the data secure for reasons of confidentiality. The data need to be encrypted so that no student-level data can be accessed and used inappropriately.

Professional Development

Several issues of professional development arose during various discussions of other components of computerized testing. Although it was not a formal topic of discussion, each presenter noted where professional development would be needed, and the participants focused on this issue as discussed in the next section of this paper.

First, teachers need to be well-versed in using computers. They need to be able to set up the computers for testing and do minor troubleshooting for many of the problems students may have. More importantly, they need to be able to integrate computers with instruction so that students are not using the computers only during assessments.

Second, teachers need professional development on how to use the scores and information produced by the computer, particularly if the computerized tests are used as interim assessments. Score interpretation can be highly dependent on test design. That is, interpreting results from a CRT administered through a CBT will be different than interpreting results from CRT administered through a CAT.

Third, a new type of training will be required for item writers. Because of the different types of item features described under a previous heading, item writers will need to be trained to consider these new features. In addition, if item writers are expected to follow a particular template, they will need to be trained on that template. In addition, appropriate training for item writers similar to that provided to those working on paper-and-pencil tests will also be needed.

District-level instructional technology staff will also need to be trained on the computerized testing system. Different training will be needed depending on whether USOE chooses a web-based system (thin client) or a local software system (thick client). If a web-based system is used, the experts recommended assembling and training a central technical support team. This team will need to be able to troubleshoot systems based on information received via a telephone call or email and be able to implement patches or correct glitches remotely. If the assessments are run onsite

through local software systems, each school will need to have access to technical support at their building. For example, there should be at least one highly trained technical support staff member in each school building or a team at the district level that work with multiple schools throughout the year. This technical support staff will also need adequate training to troubleshoot and repair any problems with the hardware, software, or interface.

Finally, the experts recommended training building and district administrators to understand the features and results of computerized testing and the implications for their building, their teachers, and their students. They will need to become both instructional and technological leaders.

SECTION II: PERSPECTIVES FROM THE SUMMIT PARTICIPANTS

The first section was intended to provide a neutral summary of the issues presented by the invited experts. This section presents the viewpoints and opinions expressed by the participants. To help gather additional information on key areas where participants' views were unclear or not specified, a meeting was held at USOE on January 22, 2008 among key USOE assessment staff, a subset of approximately 12 key Summit participants, and the two authors of this paper. For the most part, this section follows the structure of the first section, except on topics for which no opinion was expressed or on topics that were not addressed by the presenters but were raised by the participants.

Assessment Purposes and Uses

Discussion of Formative, Interim, and Summative Uses

The participants indicated a desire for interim assessments as part of a comprehensive assessment system. However, they want these interim assessments to be used for instructional purposes. Thus, the assessments must provide diagnostic information and perhaps even predictive information on how well a student is likely to do on the end-of-year assessments. Computerized assessment is appealing for interim assessments because of the fast turnaround, but the reports would need to be designed to provide maximum information. The participants also expressed a wish that the timing of the computerized tests used for interim purposes be flexible enough to fit the reality of school organization (e.g., trimesters vs. semesters).

The participants raised the issue of control over interim assessments. They felt that the interim assessment system needs to be linked appropriately with summative assessments but should provide district-level policymakers appropriate control over assessment items, timing of administration, and use of results. They did not want the results from interim assessments used in any state accountability system, and were reluctant even to allow state access to assessment results. The participants expressed a strong and unanimous viewpoint that interim assessments should be under the control of the districts, not the state. The current UTIPS system seems to be working well according to the feedback from many districts. For example, a recent survey indicated that teachers have indicated a high degree of satisfaction with UTIPS. The participants thought it was important that the summative assessments look like the interim assessments. The summit participants felt that the teachers should not have to learn different systems for the summative CBT, an interim CBT, and a formative CBT. They should all work seamlessly in the same system. That is, if USOE chooses to move to computerized testing, the participants want to use the same platform, the same software, for all types of testing. If UTIPS is run through one program, and the CRTs are administered via a very different type of program, this will hinder both the speed at which teachers will learn to use the programs, and the ability of students to use the interim assessments as practice for the summative assessments.

Additionally, the participants indicated that district leaders do not want the state to dictate curricular maps or timing. If a district teaches concept B before concept A, they want the freedom to administer an interim assessment on concept B before concept A.

Connection with Instruction

There was a strong opinion expressed that if students are using technology to take assessments, then that technology needs to be integrated into classroom instruction. The question was raised: If you only test 5% of the time, how do you effectively use computers the rest of the time? The participants expressed a desire to have computers in the classroom so that they might be better integrated into instruction. However, they recognized that it may not be feasible to have enough computers for every student to have their own. They discussed other types of technology, too, such as electronic student response systems, which are used in some districts for both instruction and assessment. Because the overall goal is a coherent system where instruction and assessment work together, the participants felt that technology may provide a pathway for meeting that goal. At the follow-up meeting, one district leader indicated that he expected USOE would need to move towards computerized summative assessment first, and then the classroom integration of technology would follow. Overall, the participants indicated that they would like to see technology support instruction first and testing second.

Focusing on CAT and CBT

The participants were interested in some aspects of CAT, but they were concerned that it would not work well in a large-scale standards based environment. And, if the CAT were to be used as an interim assessment, it would need to provide detailed diagnostic information. In general, the participants thought they could get many of the benefits of computerized testing (e.g., innovative item types, faster score turnaround, etc) through CBT and did not need to move fully towards CAT.

The second issue discussed in comparing CBT to CAT was the equating design. To develop an adequate item bank and increase the turnaround time for results, USOE would need to move toward a pre-equating system. However, there were some concerns, particularly from those with an incomplete understanding of item response theory (IRT), regarding how such a system would match Utah's goals for their students. Some expressed concern that pre-equating might be less sensitive to improvements in student learning. Participants recognized the need for shifting to a pre-equating design if they were to realize some key benefits of CBT such as instant scale score/performance level reports. They were, however, cognizant of some of the challenges necessary to move in that direction.

Overall, the districts recommended starting with CBT, because implementing a CAT may be too ambitious and the benefits uncertain at this time. However, USOE should not lose sight of the concepts of CAT, as they could be implemented in the future once everyone in the state has adjusted to a system that is 100 percent computer based.

Developing an Item Bank

The participants indicated that they felt the current item bank is good but needs to be extended from the standard and objective levels to skill levels. Participants wanted to see increased breadth and depth in item banks and developing item banks for other content areas. They were intrigued by many of the ideas presented in Dr. Parshall's framework for extending item types beyond simple multiple-choice or short-answer items. They indicated a desire to increase the relationship between items and the real world, as they thought it could both improve the measurement of the construct and increase student motivation to do their best. For instance, the participants noted that as the interactivity of the item increases, student engagement may increase and the construct can be measured at a finer grain.

The participants wanted to compare the current guidelines for developing items in Utah to the guidelines proposed by Parshall. They are open to new item types as long as there is support for using them in the classroom. Sufficient items would need to be developed to allow for the release of some of them for practice. In addition, it was noted that the type of item could drive the type of technology needed. A video-based item would have different hardware requirements than a simple multiple-choice item. The participants felt strongly that the technology needed to be installed in all districts first before complex item development could begin.

Reporting

With a move toward computerized testing, the participants indicated a wish for new types of information provided on student score reports such as distractor analysis, response time for each item, and tracking any changed answers. The participants also would like to see instant feedback provided to the student, such as right/wrong information. They felt this type of immediate feedback would increase student motivation to take responsibility for his/her own learning.

In the discussion on the current reporting structure, participants indicated that they were pleased with the timeliness of the raw score reports for the accountability assessments. They expressed a positive reaction to the UPASS reports, but indicated that they would like to see them organized in a manner that provides more information beyond summative. They also understood the value of the scaled scores and would like to receive them earlier. The participants said that they would like USOE to provide previous year's scores to teachers at the beginning of the year to inform instruction, although participants recognized that teachers would likely need additional professional development to help them use the information in the reports appropriately and learn more about data-driven decision making. Finally, they wanted to be able to integrate interim and summative assessment reports so that information from interim assessments could inform expected performance on summative assessments. However, it could be challenging to time such reports so that the information is most useable.

Infrastructure

A recent survey conducted by USOE showed that more than half of the districts reported that hardware and infrastructure are the most limiting factors for moving towards 100% computer-based testing. The discussion at the summit on infrastructure dealt largely with issues of bandwidth. The participants liked the idea of using a smart client to reduce the burden on local technical support teams while maximizing data security, as long as there was sufficient bandwidth to support many students using the system simultaneously. They also recognized the benefits of a wireless network in that it allows for easier transport of the computers and easier set up in the classrooms. However, they raised concerns about having many users on the wireless network at the same time overwhelming the system. A wireless system can require a lot of management, such as blocking rogue signals and ensuring that every classroom has a strong system. Some district leaders indicated a preference for a wired network because they felt it offers a controlled, predictable environment with hard wires. On the other hand, the concern with a wired network is having sufficient outlets for students to be able to connect to the internet. There is an issue of a tradeoff between the reliability of a wired system with a potential tangle of wires.

Another issue related to the infrastructure of the building was the availability of air conditioning. If large numbers of computers are running in a classroom, it can get quite hot, particularly during late spring through early fall. Not all buildings currently have air conditioning, and the computers may exacerbate the discomfort level of the school as well as causing the hardware to overheat.

Practical and Logistical Issues

At the same time they addressed infrastructure issues, the participants also discussed where they would want the computers to be located. Currently, many schools have computer labs with a relatively high student-to-computer ratio. For instance, several district leaders mentioned having one computer lab per school or approximately 30 computers for 500 students. Currently, these labs are used primarily for testing. The district representatives indicated that while they could accomplish computerized testing with a 30:500 computer-to-student ratio, they emphasized that a ratio of 30:300 (or 1:10) allows for both more instructional use and more flexibility to handle testing.

There was a preference for having the technology in the classroom where it could be better integrated into instruction. But they want to be able to be put it away at times as well. Laptops are preferable from a teaching perspective, but are not as durable and thus can be more costly to maintain. Likewise, the new "$100 machines" are exciting because they can get technology into every student's hands, but they limit what the students can do with them. Participants discussed other types of technology being integrated into instruction that is not computer based (e.g., whiteboards that allow teachers to print out what's been written). Many schools already have a hybrid environment with computer labs used for testing and other technology infused in the classroom for instruction. This may be a future direction for all schools to consider.

Testing windows and schedules was considered in the discussion of locating the computers in labs or classrooms. The participants expressed a concern that moving the testing window earlier in the year would reduce instructional time. Teachers prefer May testing. They also discussed the idea that if the testing window is earlier, maybe USOE could consider testing on a calendar year (e.g., test material taught from February 2006 to February 2007). However, the participants liked what they were hearing from the experts; namely that computerized testing should reduce the turnaround time for scoring and allow us to move the testing window as late in the year as feasible.

Professional Development

If Utah moves to computerized testing, then teachers need to learn how best to integrate technology into instruction. However, the participants were clear that they would prefer that the driver be instruction rather than assessment. That is, teachers should first work to integrate technology into instruction and then computerized testing should follow. To support professional development in technology, the participants recommended using a model similar to how the state has approached improving literacy instruction with state supported literacy teams and coaches. The participants argued that the state should consider creating and supporting technology teams and coaches. Ideally, participants would like to see each school have a technology expert with either content expertise or to pair a technology person with a content person. Coaches should be strong in technology and one content area, and this should be a funded position.

In addition to incorporating technology into instruction, teachers will also need to know how to interpret the results from the computer-based assessments. All teachers need training on data-driven decision making. One example cited by the district representatives is some frustration among teachers about how to combine and distill information from multiple sources. That is, if Dibels, UTIPS, and the CRTs provide different information about a student, how can a teacher integrate that information to determine how best to address the student's needs? The participants recommended that teachers receive professional development in context by working with experts using real data. These types of training sessions will help teachers understand what they can learn from their existing data and how to use it to determine next steps. State level professional development programs would help with consistency across districts and prevent each district from reinventing the wheel. Although integrating technology into instruction is currently happening at schools without state-level training, some state guidance on best practices could help more schools achieve this goal more rapidly.

Finally, teachers will need access to technical support both during assessment and for on-going classroom instruction using technology. Approximately 41% of the respondents to the USOE survey indicated that a lack of technical support is one of the top two most limiting factors for implementing computerized testing. Currently, the average number of FTE's allocated for technology support per school is approximately one-third of a full-time staff member. Ideally, the district leaders indicated they would like to see technical support provided at multiple levels, such as a toll-free number,

district-level technical support, and building-level support. It will be important to understand who owns which problem and how to diagnose the technology problems.

Other Policy issues, concerns, and recommendations

When asked about other policy issues, concerns, or recommendations, the participants raised two issues not previously discussed: procurement and funding.

Procurement

In the discussion on procurement issues, two aspects were discussed: hiring vendors to support the development of computerized tests and buying the necessary hardware and equipment.

In the discussion regarding vendors, the participants indicated that district policymakers want the state to enter into contracts longer than the typical three agreements. They stated a belief that a long-term relationship will help to ensure that the vendor is vested in the success of the project. One idea discussed was to bid a five-year contract with possible year-long extensions with evaluations attached to them. That would provide vendors with the incentive to stay fresh, while allowing them to build relationships with the school districts.

There was also a fairly heated discussion regarding the purchase of equipment. Districts and schools want more flexibility in choosing what to purchase. They do not want to be sent computers or told which computers to buy. The cheapest is not always the best. For instance, laptops are invariably more expensive than desktops—both in terms of purchase price and upkeep cost—but, in some situations, laptops are easier to integrate into the curriculum.

Funding

The participants would like to see a minimum computer-to-student ratio funded by the state for each school. They want a hard budget for buying and maintaining both hardware and software. The participants felt that more attention was paid to start up costs than to maintenance costs. The district representatives appreciated the $50 million influx of state money this past year, but they were very worried about the stability of an infrastructure funding stream. They expressed a concern that many districts could not commit to supporting the infrastructure necessary to ensure 100% computerized testing implementation without some strong sense that state funding would continue to help support this infrastructure. They need to have flexibility to pay for ongoing maintenance. They did not want the state to dictate that funds could only be used for the purchase of new equipment if what was really needed was repairs or upgrades on the current equipment. They also need ongoing resources to fund the training and professional development for both teachers and technical support personnel. In addition, the districts want to be able to plan more than one year into the future and know that funding will be available to support that plan. A strong conclusion expressed was that districts need stability and predictability in their technology budgets.

Summary and Next Steps

The summit was a good start for USOE to begin considering the issues associated with moving to computerized testing. The districts were supportive of moving the summative assessments (CRTs) to computer-based testing as long as basic infrastructure, staffing, and training needs were met. Overall, a vision emerged of a coordinated approach to assessment in Utah and one which allowed technology to be integrated into instruction so that computerized testing would tie seamlessly to instruction. That is, participants felt that administering a test via computer was only a small part of the larger picture. Districts wanted a cohesive structure for instruction, formative assessment, and interim assessment. Then, the summative assessment should follow from that.

However, to achieve this vision, additional questions must still be addressed. The following is a list of questions for USOE to consider as the state moves forward towards adopting a computerized testing program. The questions should be considered as a set because the answers to one question will likely have direct implications for other questions. One consideration the committee put forth was to recognize that districts are currently in different places in terms of their current infrastructure and implementation of technology and face different obstacles to implementation. Therefore, there will not be a standard answer to most of these questions that will suit all districts. However, there should be philosophical underpinnings to the statewide vision of computerized testing so that all districts and schools are moving in the same direction.

1. Should computerized testing be used for accountability tests? Interim assessments? Both?
2. How will the existing computerized interim assessments (UTIPS) work with new computerized accountability tests to produce a balanced system?
3. Does USOE want to pursue a CBT or CAT system?
4. Does USOE want to add new types of items that are possible only in computerized testing systems? If so, how will the item writers and reviewers be trained on these new item types?
5. What are the implications of moving toward a pre-equating system to increase turnaround time for results?
6. How many computers will need to be purchased?
7. Is it more beneficial to consider desktop or laptop computers? Will each district be allowed to make their own decisions based on their current status?
8. Should the computers be housed in classrooms, a lab, or on mobile carts? Will the answer vary by district depending on the age and layout of their buildings?

9. Should the computers be connected to the internet through wires or with wireless technology? Can different answers be given to different districts?

10. Should the assessment be administered using a thin, thick, or smart client?

11. What type of technology support is needed to minimize any technological problems with computerized testing?

12. What type of professional development is needed to help teachers integrate technology with instruction appropriately?

13. How will the state meet the demand for on-going resources to support the use of technology both in instruction and assessment?

14. How will the technology be adapted to meet the needs of special interest students, such as English language learners and students with disabilities?

APPENDIX A: SUMMIT AGENDA

Utah State Office of Education

Computer-Based Testing (CBT) Summit

November 7-8, 2007

Location: Best Western Hotel

154 West, 600 South, Salt Lake City

Tuesday, November 7

8:00 Continental Breakfast

8:30: Welcome

- Superintendent Patty Harrington

8:45 Goals of the Summit and Setting the Context

- Deborah Swensen, Assessment Director
  - ✓ Legislative and State Board requirements
  - ✓ Current levels of CBT implementation in Utah
  - ✓ Current readiness levels—the state of software, hardware and infrastructure

9:30 Plenary Panel: A Vision for Online Testing. Panel members will be asked to outline a forward-thinking vision, but one that should be achievable in 5-10 years (15 minutes each)

- Content
  - ✓ Item delivery (presentation issues) and item types
  - ✓ CAT and/or CBT
  - ✓ Equating issues
- Access and accommodations
  - ✓ Comparing access issues between CBT and Paper/Pencil
  - ✓ Type of accommodations available via CBT
- Scoring, Reporting & Integration
  - ✓ Scoring selected- and open-response items
  - ✓ Report formats and timing
  - ✓ Integration with Student Information Systems (SIS) and Data Warehouse
  - ✓ Integration with interim and formative assessment
- Infrastructure
  - ✓ Hardware
  - ✓ Software
  - ✓ Bandwidth

- ✓ Platform
- ✓ Estimating costs for hardware, software, and connectivity

- Pragmatics and Logistics
  - ✓ Administration and timing
  - ✓ Space and housing
  - ✓ Challenges in large and small school districts

10:45 Break

11:00 Brief Q & A with panelists

11:30 In-depth presentations

- Content
- Infrastructure

The expert who presented the overview of each of these particular topics will present an in-depth look at the issue related to their area of expertise. This will be a 30 minute presentation followed by 60 minutes (after lunch) of interaction to help articulate the long-term vision for computer-based testing in Utah.

12:00 Lunch

1:00 Continuation of morning in-depth presentations

- Content
- Infrastructure

One hour of interaction to help articulate the long-term vision for computer-based testing in Utah.

2:00 Report out from workgroups

2:30 Break

2:45 In-depth presentations

- Access and accommodations
- Reporting and Integration

The expert who presented the overview of each of these particular topics will present an in-depth look at the issue related to their area of expertise. This will be a 30 minute presentation followed by 60 minutes of interaction to help articulate the long-term vision for computer-based testing in Utah.

4:15 Report out from 2:45 workgroups

4:45 Adjourn

Wednesday, November 8

8:00 Continental Breakfast

8:30: Reconvene

- Questions and review of key issues from Day 1

8:45 In-depth panel discussion about pragmatics and logistics

This panel will include the expert who presented the overview as well as representatives from Utah's large and small school districts. Each of the three panelists will present for approximately 15 minutes. This will be followed by an hour of interaction and discussion to frame the vision. This session is designed to frame the issues associated with different practical and logistical implementations issues faced by large schools and districts compared with small schools and districts.

10:30 Break

10:45 Report out and debrief

11:00 How do we get there from here?

- Building a policy framework to support the long-term vision
- Overcoming obstacles to the vision
- Learning from other states successes and challenges
- Learning from full implementation districts
- What are reasonable short-term, intermediate, and long-term timelines?

12:45 Next steps

- "White paper"
- RFP

1:00 Lunch

APPENDIX B: ACCOMMODATIONS AND COMPUTER-BASED TESTING

From Thompson, Thurlow, Quenemoen, and Lehr (2002)[8]

The following tables from Thompson, Thurlow, Quenemoen, and Lehr (2002) list various accommodations to address the needs of students with a variety of accommodation needs—including students with disabilities, students with limited English proficiency, students with both disabilities and limited English proficiency, and students who do not receive special services, but have a variety of unique learning and response styles and needs. For each accommodation, relevant considerations are provided in the tables. The three columns to the right of the Considerations Column represent:

- A built-in feature of universally designed computer-based tests (available for self-selection by any student)
- The need for this accommodation is not affected by computer-based testing
- A new or different accommodation may be needed for computer-based testing

Table 5. Presentation Accommodations

Accommodation	Considerations for Computer-based Tests	1* Built in	2* Not affected	3* Other needed
Large print and magnification	Capacity for any student to self-select print size or magnification Graphics and text-based user interfaces have different challenges Scrolling issues Determination of optimal print size (e.g., default set at 14 pt) and size of graphics to reduce need for large print or magnification Variations in screen size Effects of magnification on graphics and tables	X		X

[8] Excerpted from Thompson, S. J., Thurlow, M. L., Quenemoen, R. F., & Lehr, C. A. (2002). *Access to computer-based testing for students with disabilities* (Synthesis Report 45). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved November 1, 2007, from the World Wide Web: http://education.umn.edu/NCEO/OnlinePubs/Synthesis45.html

Instructions simplified/clarified	Instructions designed for maximum simplicity/clarity Capacity for student to self-select alternate versions of instructions in written or audio format Capacity to have instructions repeated as often as student chooses Variable audio speed	X		
Audio presentation of instructions and test items	Capacity for any student to self-select audio (screen reader) presentation of instructions (all students wear ear/headphones) Graphics and text-based user interfaces have different challenges Capacity to repeat instructions and items as often as student chooses Variable audio speed Audio presentation must be high quality	X		
Instructions and test items presented in sign language	Capacity for student to self-select alternate versions of instructions in written format Capacity for student to self-select signed versions of instructions and test items (Note: some words may not be easily translated into sign language) Graphics and text-based user interfaces have different challenges Not feasible to read lips on video	X		
Instructions and test items presented in a language other than English	Capacity for student to self-select alternate language versions of test items in written or audio format Beware of the speed at which some languages are produced, they may take more space than English Machine translation capability Graphics and text-based user interfaces have different challenges Capacity for pop-up translation Variable audio speed	X	X	

Braille	Use of screen reader that converts text into synthesized speech or Braille Alternative tags for images Graphics and text-based user interfaces have different challenges Students using Braille may require extra time		X	X
Highlighter and place holding templates	Capacity for any student to self-select highlighter Graphics and text-based user interfaces have different challenges Clear instructions for use of highlighter	X		
Graphics or images that supplement text	Careful selection of images Alternative text or "alt tags" for images Graphics and text-based user interfaces have different challenges Avoidance of complex backgrounds or wallpaper that may interfere with the readability of overlying text Tactile graphics or three-dimensional models may be needed for images			X
Paper/pencil test format	Students who are not computer literate Students who need accommodations that are not available on computer-based assessments			X
Use of color	Capacity for multiple choices of screen and text color Graphics and text-based user interfaces have different challenges	X		
Flashing or blinking text or objects	Flash or blink frequency	X		
Multiple column layout	Linear presentation order needs to be logical Access to screen reader Graphics and text-based user interfaces have different challenges	X		X
Captioning	Capacity to provide captioning for video content	X		X

Table 6. Response Accommodations

Accommodation	Considerations for Computer-based Tests	1* Built in	2* Not affected	3* Other . needed
Write in test booklet	Capacity for multiple options for selecting response – mouse click, keyboard, touch screen, speech recognition, assistive devices to access the keyboard (e.g., mouth stick or head wand) Speech recognition capability	X		X
Scribe	Capacity for multiple options for selecting response (see above) **Speech recognition capability	X		X
Brailler	Speech recognition capability		X	X
Tape recorder	Speech recognition capability			X
Paper/pencil response	Student computer literacy Option for paper/pencil in addition to computer (e.g., scratch paper for solving problems, drafting ideas) Option for paper/pencil in place of computer (e.g., extended response items) Speech recognition capability		X	X
Spell check	Capacity for any student to self-select spell check option Capacity to disable spell check option when spelling achievement is being measured Spelling implications when using speech recognition software	X		
Calculator	Capacity for any student to select calculator option	X		
English or bilingual dictionary/glossary	Capacity for any student to self-select dictionary option Capacity for Pop-up definitions of key words/phrases (built into assessment) Capacity for use of multiple languages	X	X	X

Table 7. Timing/Scheduling Accommodations

Accommodation	Considerations for Computer-based Tests	1* Built in	2* Not affected	3* Other accom. needed
Extended time	Availability/location of computers and peripherals Flexible, individualized timing	X	X	
Time of day beneficial to student	Capacity of network system Availability/location of computers and peripherals Scheduling	X	X	
Breaks during test	Maintaining place and saving completed responses Capacity to turn off monitor/ blank screen temporarily Test security		X	
Multiple test sessions, possibly over multiple days	Maintaining place and saving completed responses Test security		X	
Order of sub-test administration	Capacity of technology for self-selection of subtest order Test security Scheduling	X	X	

Table 8. Setting Accommodations

Accommodation	Considerations for Computer-based Tests	1* Built in	2* Not affected	3* Other accom. needed
Individual or small group administration	Availability/location of computers and peripherals Grouping arrangements Use of earphones or headphones Use of individual setting if response method distracts other students (e.g., speech recognition)	X	X	
Preferential seating	Availability/location of computers and peripherals	X	X	
Special lighting	Minimize glare from windows or overhead lights Contrast and color in item design	X	X	
Adaptive or special furniture	Types of adaptive furniture needed and need for changes in computer location/ peripherals		X	X
Hospital/home/non-school administration	Availability/comparability/location of computers and peripherals Test security		X	