

# **Use of Technology-Supported Tools for Large-Scale Science Assessment: Implications for Assessment Practice and Policy at the State Level**

**Edys S. Quellmalz and Geneva D. Haertel  
Center for Technology in Learning  
SRI International**

Paper commissioned by the Committee on Test Design for K-12 Science Achievement  
Center for Education  
National Research Council

*Copyright © 2004 National Academy of Sciences. All rights reserved. No part of these pages, either text or image may be used for any purpose other than personal use. Reproduction, modification, storage in a retrieval system or retransmission, in any form or by any means, electronic, mechanical or otherwise, for reasons other than personal use, is strictly prohibited without prior written permission.*

*Opinions and statements included in the draft papers are solely those of the individual author(s), and are not necessarily adopted or endorsed or verified as accurate by the Committee on Test Design for K-12 Science Achievement or the National Academy of Sciences, including the National Academy of Engineering, Institute of Medicine, or National Research Council.*

A half century ago, the testing industry was revolutionized by the use of multiple-choice item formats in combination with machine scoring technology. With these advances, educational institutions were able to capture evidence of student learning in a cost-efficient and standardized manner and return the information quickly to school districts nationwide. Today, educators and test publishers are in the midst of a second technology-supported revolution that has been catalyzed by the advent of computers, video, and network technologies. This second revolution will be more powerful and pervasive than the first.

In the late 1990s, the Panel on Educational Technology of the President's Committee of Advisors on Science and Technology (1997, p.33) acknowledged this possibility as follows: "the real promise of technology in education lies in its potential to facilitate fundamental, qualitative changes in the nature of teaching and learning." Since then, leaders in the assessment industry have further elaborated this vision indicating that technology will push large-scale testing well beyond the familiar paper-and-pencil formats to measure high-level cognitive processes and problem-solving strategies that heretofore were out of reach. They predict that technology will not only transform the types of performances we will be able to measure, but will also transform the way assessments are designed, developed, administered, scored, and reported (Bejar, 1996; Bennett, 1998; 2001;2002; Mislavy, Almond, Steinberg, Haertel & Penuel, 2003;). Bennett (1998, p.10) provides a glimpse of the predicted technology-based transformation of testing,

In sum, several key transformations will define this next generation of large-scale tests. These transformations will be in the character of questions, development and scoring processes, test design, and test center networks. Most notably, the ability to deliver multimedia questions, capture and score complex constructed responses, create tasks efficiently, and move and manipulate large amounts of data electronically will make performance assessment a vital, if not principal element of large-scale testing.

This paper was commissioned by the National Research Council to provide guidance to states on ways that technologies can be harnessed to support development of sound science assessment systems. In this paper, we will examine the current state-of-practice in technology-supported assessment, with examples drawn from several content areas, but with an emphasis on the content domain of science, including inquiry skills. Our goal is to apply our conclusions about technology-supported assessment to improving practice and policy for state science assessment. In this paper, we:

- Briefly review the status of current large-scale student assessment;
- Overview the state-of- practice in technology-supported assessment;
- Review current and near term technology supports for the implementation and development of assessments and assessment systems;
- Describe technology supports for designing articulated assessment systems;
- Describe the potential for the assessment of complex science learning;

- Index technological advances to their assessment roles, including their application to components of science inquiry;
- Offer scenarios of current, near-term and innovative ways that states can employ technology to transform their science assessment systems;
- Conclude with a review of the challenges and promise of technology supports for assessment

### **Status of Standards-Based Student Assessments**

The recent passage of the No Child Left Behind Act has increased pressures on state departments of education to develop more effective and efficient strategies for measuring student performance. The use of technology-supported assessment is one such strategy. Although reforms are taking place in the setting of standards and design of professional development to support better teaching, learning, and assessment, appropriate forms of large-scale assessment are in short supply. Recent evaluations of the statewide system initiatives revealed that states are still relying primarily on traditional student outcome measures that are not tightly aligned with reform goals (Lawrenz & Huffman, in press; Klein, Hamilton, McCaffrey, Stecher, Robyn, & Burroughs, 2002; Webb, Kaufman, Kaufman & Yang, 2001). One of the most frequently voiced criticisms is that existing standardized, multiple-choice tests have limited capacity to measure critical-thinking and problem-solving, especially in science inquiry where higher-level problem-solving is crucial.

**Limitations of Current Paper-Pencil Student Assessments.** Traditional, on-demand, paper-pencil tests still favor breadth of content coverage over depth of reasoning, and it is generally acknowledged that most large-scale assessments are not aligned with current goals to increase student learning of complex, conceptual understandings and high-level problem-solving and inquiry skills. Large-scale assessments tend not to tap the expanded repertoire of student outcomes judged to be important in the cognitive psychology literature (Pellegrino, Chudowsky, & Glaser, 2001) including deep subject matter understanding, inquiry strategies, communication, metacognitive/self monitoring strategies, and collaboration. Few large-scale assessments, even performance assessments, probe in detail the fundamental ways that individuals process and use information in tasks that require extended lines of reasoning (Baxter, Elder, & Glaser, 1996; Quellmalz, 1984). Moreover, as states and districts invest heavily in technology, they seek measures of the impacts of technologies on student learning and the progress of their students in the fourth literacy: technology literacy. Thus, technology offers students alternative methods for engaging in assessment tasks that would reveal a greater range of cognitive skills and processes.

The limitations of most currently available science and mathematics assessments in measuring these higher-level cognitive processes and skills are highlighted in analyses revealing their weak alignments with national standards developed by professional organizations, such as the *National Science Education Standards* (NSES) (National Research Council, 1996), the *Benchmarks for Science Literacy* (American Association for the Advancement of Science, 1993), and National Council of Teachers of Mathematics (NCTM) *Principles and Standards for School Mathematics* (NCTM, 2000). In a study funded by the National Science Foundation, “Validities of Science Inquiry Assessments”, Kriekemeier, Quellmalz, and Haydel (2004) found that the percentage of items judged to test science inquiry skills in three reference exams, the Trends in Math and

Science Survey (TIMSS), the National Assessment of Educational Progress (NAEP), and the New Standards Science Reference Exam (NSSRE) was very uneven, ranging from 35% to 72%, with a number of inquiry standards such as posing questions and communication not addressed at all. A related cognitive analysis study of students thinking aloud as they responded to questions from the three exams found that performance assessments elicited more inquiry skills than items in the constructed-response or multiple-choice formats (Quellmalz & Haydel, 2002). Even for the open-ended items, students were able to provide more evidence of their inquiry skills than formal test questions elicited. More importantly, the study lends empirical support to claims that many science inquiry standards, such as formulating scientific explanations or communicating scientific understanding, are not measured by using a multiple-choice format with one correct answer.

Nor, can many significant, complex problems of science be presented well, or at all, in print format. In the physical sciences, for example, students' understanding of the ways that heat, light, mechanical motion, and electricity behave during nuclear reactions could be readily assessed in a computer simulation, but depicting the complexity of such dynamic interactions is too difficult for a print, multiple-choice format and too dangerous for hands-on experiments. Likewise, in the life sciences, understanding the multiple ways that cells interact to sustain life, or the passage of hereditary information through processes such as meiosis, gene alterations, or crossover, can be depicted best by using dynamic models that demonstrate the effects of changes at one biological level of life (e.g., DNA, chromosome) on other levels of life. Technology-based assessment tasks that take advantage of simulations developed for curriculum programs are well suited to assess knowledge of these highly interrelated systems. Similarly, knowledge of ecosystems, the effects of the rock cycle, patterns of atmospheric movement, and the impacts of natural and biological hazards on species can be more appropriately measured by using technology-based assessments. Such assessments can represent natural or man-made phenomena, systems, substances, or tools that are too large, too small, too dynamic, too complex, or too dangerous to be adequately represented in a paper-and-pencil test or a performance test format. Bennett (2001, p.8) quotes the report to the President and Congress of the bipartisan, Web-Based Education Commission (Kerrey & Isakson, 2000) with regard to the promise of technology for assessment practice:

Perhaps the greatest barrier to innovative teaching is assessment that measures yesterday's learning goals . . . Too often today's tests measure yesterday's skills with yesterday's testing technologies—paper and pencil. (p.59).

**The Promise of Technology.** Technology-supported assessments can overcome the on-demand, one-shot, decontextualized problems typically associated with traditional multiple-choice achievement tests. In addition to permitting assessment of a broader range of phenomena, technology-supported assessments can be designed to incorporate the more tractable features of alternative forms of assessment. Technology-supports can make the advantages of these alternative assessment forms available, provide more cost-efficient methods for scoring and storage of artifacts, and make the results more quickly available to stakeholders. The most innovative forms of these technology-based assessment methods are currently implemented primarily in technology-rich science curriculum programs, however researchers are exploring

ways to extract and design science assessment tasks that can be administered and scored in systematic ways that will produce evidence of technical quality needed for accountability purposes.

These types of advances in technology-based assessment will permit educators to keep pace with the kinds of knowledge and skills that students are required to learn in the 21<sup>st</sup> century. Technology can facilitate the instantiation of cognitive principles in the design, delivery, and scoring of assessment tasks. Assessments that are technology-supported will present new kinds of stimulus environments, ask students different kinds of questions, use a greater range of response formats, capture student responses easily, and support the scoring, archiving, and reporting of the results of these qualitatively different types of assessments. Below we describe how states are beginning to take advantage of technology to strengthen and transform their science assessment systems. In addition, we describe technology applications that can significantly improve and ultimately transform state science assessment systems.

### **Technology Supports for State Assessment Operations**

States' entry into the realm of computer-based testing has tended to begin with supports for testing operations. States look to electronic testing for cheaper and more efficient test delivery, more timely return of scores, ease of data analysis, and flexibility of report generation for a variety of audiences. Concern for speed and economy have led to an emphasis on multiple-choice formats that permit automated scoring. A growing trend is a class of computer-based test products marketed to districts and schools as instructional resources to prepare students for the state tests. The science education community may view such computerized test preparation on traditional test formats as incompatible with the forms of assessment needed to measure achievement for challenging science standards and curricula.

In this section, we consider how methods involved in the implementation of large-scale assessments can take advantage of the affordances of technology. We describe uses of currently available technological capabilities, innovative, but near-term efforts, and promising, longer range possibilities.

**Online Delivery.** The number of states venturing into computer-based testing is on the rise. Increased testing requirements and shrinking budgets are driving states to find less expensive, more efficient testing methods. *Technology Counts* (Education Week, 2003) described how twelve states and the District of Columbia are administering computer-based assessments. Notably, only four of those states were administering science tests online, with only one state including open-ended responses. In 2004, Maine moved its innovative multi-format science assessment to laptop delivery.

One of the earliest states to convert to technology-supported delivery was Oregon as part of its Technology-Enhanced Student Assessment (TESA) project. Neuberger (2004) reported that the cost of developing the online reporting system was recovered within one year. Commercial publishers assert that the actual costs of putting a test online is comparable to developing a good paper-pencil exam. Vendors estimate that computer-administered tests save half to three quarters of the administrative costs of paper-pencil versions.

Oregon’s empirical studies of the technical qualities of the print and online versions of the assessments yielded comparable item statistics. Testing companies also report conducting comparability studies with similar findings, although these studies are not in the published literature. Russell, Goldberg, and O’Connor (2003) have suggested that the comparability of standards-based assessment results, that use different presentation modalities, such as print, face-to-face instructions, or computer-based delivery, may be less difficult to attain than establishing that a student has had the “opportunity-to-learn” content that is delivered by computer. Others, however, point out that once electronic assessments take advantage of technologies to present types of complex problems not practicable in conventional formats, the comparability question will not apply.

As long as states require accountability testing to occur within limited testing windows, the issue of infrastructure availability and capacity will pose significant obstacles to large scale administrations. The capacity of technology to support shifts to periodic, just-in-time administration holds promise for both alleviating system load and making results more timely and instructionally useful.

**Adaptive Testing/Scaffolding.** The capacity of computers to tailor items to students’ ability levels is attractive both for the efficiency of large scale testing and the diagnostic utility of formative assessment. Adaptive testing may require the use of a very large and cross-classified item bank that can select items for multiple targets and provide rapid feedback based on items with varied levels of scaffolding. In an adaptive assessment, a technology-based tool can be used to construct the item pool from which the assessment will be assembled. Almond, Steinberg, and Mislevy (2002) point out that there are two key processes involved in identifying items to be included in an adaptive assessment—a selection process that determines whether or not to present a given task; and a sequencing process—in which the order in which the item or task to be presented is determined. The technology-based tool must know where to start the assessment and where to end it. The technology-based tool has to monitor information about the test taker’s knowledge, skill, and abilities and then update the test taker’s proficiency on these attributes after each item is answered in order to determine what item or task should be presented next. Depending on how many different types of knowledge and skills are being assessed in the test, the tool may be required to monitor an overall proficiency or several different proficiencies.

When an adaptive assessment is being used for diagnostic purposes, the technology-based tool must be designed to select an assessment task with appropriate levels of scaffolding tailored to a test taker’s skill level. Thus, the tool must be designed to provide test items or tasks that have scaffolds to varying degrees. Items with scaffolds include more supporting information in their stimulus and response materials and lessen the load of information that a test taker brings to the performance.

Several test publishers offer adaptive testing products that states and districts have used. But, because adaptive testing has been interpreted as out-of-level testing, the use of it for NCLB requirements has been denied. States such as Oregon and South Dakota have deferred use of adaptive testing or shifted its use to low stakes, district testing programs.

In science assessment, the Diagnoser, developed by Minstrell (1995) as an adaptive testing tool for physics that provides several levels of items with scaffolds, is being employed in curriculum program and district-level assessments.

Other, innovative examples of adaptive testing environments have been developed in the content domain of science. For example, the BioMASS Project (Steinberg, Mislevy, Baird, et al., 2003), a high school biology curriculum for use in Advanced Placement classes, makes use of technology-supported simulations in two topical units. In one unit, students engage in simulations to discover dominance patterns of three physical characteristics of mice. In the second simulation, which focuses on micro-evolution, they engage in a field-experiment to study two populations of lizards that live in environmentally diverse conditions. The simulations support students' testing of hypotheses, planning of a field experiment, and interpreting results from scientific studies.

In the Integrative Performance Assessments in Technology model, prototype assessments were developed using a modular design to tailor the administration of different assessment tasks to students depending on their familiarity with the embedded technology tools and the complexity of science and technology required (Quellmalz & Zalles, 2002a; 2002b; Quellmalz & Kozma, 2003; Means, Penuel, & Quellmalz, 2002). These projects provide examples of ways technologies can support the assessment of complex, extended science inquiry within problem-based scenarios.

**Automatic Scoring.** The technology supporting the scoring of responses has been rapidly evolving and reflects advances in semantic analysis and computer-based scoring of written text. Typically, online scoring is done automatically for multiple-choice items, while short constructed-response questions are scored by human raters, either online or in face-to-face sessions. Shermis (2004) reports increased comparability of essays of over 100 words scored by computers and by human scorers. A number of commercial products support automated essay scoring. Consistent methods for shorter constructed-responses are still being refined.

States such as Oregon and Pennsylvania have used essay grading programs for their statewide writing assessments. Indiana has piloted automated scoring for end-of-course exams. The approaching automation of short constructed-response scoring should catapult testing of student thinking and reasoning in science and other subjects to a new level.

**Online Rater Training and Scoring.** In this activity, assessors transmit some types of complex, open-ended responses to human graders online which eliminates the need to transport the results of the examinations, store the paper-pencil results, and provide hospitality to scorers at a central location (Odendahl, 1999; Whalen & Bejar, 1998). Technology reduces the logistics and costs associated with scoring open-ended items. Computer supports for live scoring have been highly developed by commercial testing companies. The supports range from online training and interspersed calibration checks for raters scoring paper-pencil essays in distributed locations, to fully online systems where student work is scanned. States and test publishers have also developed online training and practice sets to acquaint teachers with scoring criteria so they will use comparable ones in their classroom assessments.

Online guides have been created to support the provision of formal rater training and scoring, informal practice scoring sessions, and interpretation and instructional diagnosis based on test performance. The Maine Assessment Portfolio (MAP) provides an online professional community that supports practice scoring of student work. (See [http://www.maptasks.org/pr\\_.html](http://www.maptasks.org/pr_.html)) Teachers are provided with scoring practice using analytic, task-specific rubrics and scoring guides that are associated with each of the assessment tasks. Samples of student work that have been scored by Maine teachers are available online. Those teachers who are learning how to score results are encouraged to assign their own scores to the student work samples and get instant feedback on their inter-rater agreement. Online training exercises focus on improving scoring accuracy, consistency, and identifying particular kinds of evidence needed to meet the state of Maine's content standards as defined by each MAP task.

Another example of online rater training and scoring is available on the Performance Assessment Links in Science Website (Quellmalz & Schank, 1998;). (See <http://www.pals.sri.com/guide/scoringlearn.html>). This online rater training permits geographically distributed individuals or groups to learn to apply rubrics reliably to the science performance assessments that appear on the PALS Web site. Training and score entry can be completed online and student responses may be digitized or paper-pencil.

The use of online shared workspaces can support assessment scorers to participate in a group training online. Microsoft's Net Meeting is an example of communication and presentation software that permits participants to receive identical training where all scorers can see the same examples, interact with other meeting participants, share comments on the student work samples and rubrics, and amend the scoring rubrics as a group, if necessary. Using an online system, raters can enter their scores of student work online and their agreement with scoring standards can be monitored and rater drift can be periodically calculated.

**Analysis and Reporting of Assessment Scores.** A large-scale assessment system can include both formative and summative items and tasks. For the purposes of this paper, we will proceed as if we are dealing with an articulated state assessment system that has assessments at the state, district, and classroom levels and that results would be reported for each of these levels. In such a system, it is likely that a state would disaggregate the data it collects to present average scores by different subgroups that might represent regional areas of the state, counties, school districts and schools. Individual student data would likely be disaggregated by gender, race/ethnicity, economic status, English proficiency, and grade. The results for particular subgroups could be further described by providing data on the percent of students with primary disabilities in need of special education and the percent of limited English proficiency students.

The formative and summative items that would be administered to classrooms as part of an integrated large-scale assessment system would be aligned with the range of enacted curricula. The formative items would be used to make instructional decisions and the summative items would be used to make judgments about accountability. The summative items would have to demonstrate psychometric properties such they could be used by state, district and school administrators to make policy and programmatic decisions. These items should display levels of technical quality that are set forth in the AERA/APA/NCME (1999) guidelines. New summative

items would have to be developed, pilot tested, and field-tested periodically and new IRT parameters calculated.

In order to document the psychometric features and technical qualities of their assessment programs, states may work with an advisory committee that includes psychometric expertise. Current psychometric models and methods are available that describe student achievement in terms of measuring multiple proficiencies, chart student progress over time, identify the qualitative level of performance that an individual has reached during a complex performance; and provide a variety of qualitative and quantitative feedback at the student, group, class, school, and state level. The type of psychometric information that a state reports depends on the nature of the inferences that educators, parents and policymakers want to make from their assessment data, the ways that learning and schooling are understood, and the technologies that are used to gather and display the test data.

Psychometric tools are currently available to conduct the analyses needed to support large-scale assessment programs. The use of technology, in the form of statistical software programs to support psychometric analyses, has evolved such that the four general classes of measurement models—Classical Test Theory, Generalizability Theory, item response modeling and latent class—are available. Software is available to calculate psychometric information that ranges from reliability indices, standard error estimation, test equating, generalizability studies, modeling of unidimensional and multidimensional item responses, and models of change and growth, including the incorporation of cognitive elements in these models. Also, there have been advances in the psychometric modeling of cognitive structures using such techniques as differential item functioning, hierarchization, Bayes Nets, and the use of the unified model and M<sup>2</sup>RCML. (See Pellegrino, Chudowsky, and Glaser (2001) for a discussion of these various psychometric models and tools.) While it is beyond the scope of this paper to specify all of the psychometric software available, Bilog, Multilog, ConQuest, and GradeMap, exemplify the types of specialized psychometric packages that can be used to conduct item modeling analyses of data. Other more general statistical software packages, such as SAS, SPSS, LISREL and M-plus can be used for the purpose of calculating generalizability and validity studies.

There are several levels of reporting that are needed to communicate large-scale assessment results. Teachers need to understand where their students stand with regard to the assessment targets and what they will need to do instructionally to help their students achieve proficiency. School administrators need to understand where their school stands and what instructional programs need to be implemented to improve performance. District- and state-level policymakers need to know where the schools under their leadership stand and what can be done to help each of them attain accountability goals. In states that use Web-based systems, it is possible to make the following types of information available to stakeholders:

- the percentage of individual and subgroups of students at each grade level who attained each standard;
- the percentage of students at a particular grade who attained each standard compared to the percentage of students who attained the standard at that grade level in prior years;
- the percentage of students who attained the standard at each grade level in a particular school with their counterparts from other schools in the state and to the state, as a whole;

- the percentage of students who attained the standard at each grade level in a particular district with their counterparts from other districts in the state and to the state, as a whole;

Several states have created Web-based systems to assist these different groups of stakeholders in understanding and using assessment data (Wayman & Yakimowski, 2004). One example is the North Central Regional Educational Laboratory's (NCREL) WINSS Successful School Guide (<http://www.dpi.state.wi.us/sig/index.html>). The WINSS Web site provides Wisconsin school personnel, parents, and community members with information on the standards, assessment, data analysis and best practices. Users of the site are able to track information about their local schools and districts and compare their performances with other relevant schools and districts. Two tools that are tailored to the needs of particular stakeholders include Maryland's data analysis tool and the Maryland online staff development course to help principals understand the instructional targets of Maryland's accountability system and how the assessment data can inform planning for school improvement (<http://mdk12.org/data/index.html>). These two sites illustrate the types of sites that are being developed nationwide to provide technical assistance and tailored reporting of assessment data. Online resources are an effective means of distributing reports to stakeholders while cutting the costs of printing and mailing hardcopy.

### **Technology Support for State Assessment Development**

In the following sections, we describe the various components and functions associated with the development of a large-scale assessment system, that could involve multiple, articulated levels (i.e., classroom, district, state) and how technology tools can contribute to more rapid development, higher quality assessments, greater efficiency and lower costs as compared to conventional paper-pencil, large-scale systems.

**Alignment with Standards.** States and districts are motivated to align their standards to the national standards in subject areas. Technology can facilitate such alignment activities by offering online procedures and formats for creating databases that link a state's standards to national and district standards. Alignments can be incorporated in relational databases that can be searched using sophisticated strategies and dynamically identify assessments that are judged to test specified standards.

One type of technology-supported alignment provides information about the degree of alignment between a set of content standards and student assessments. Experts at distributed locations receive "alignment literacy" training on an online system and enter their judgments of how aligned the standards are with the assessment using the Web-based tool. Norman Webb at the Wisconsin Center for Educational Research (WCER) has developed such a tool (Webb, 1999).

The Performance Assessment Links in Science (PALS) database, funded by the National Science Foundation (NSF), pioneered the design of digital libraries of hard-to-find, expensive-to-develop science performance assessments as well as alignment tools to support searching for the assessments by standards or curriculum units (Quellmalz, 2003) (See <http://pals.sri.com>). Methods derived from the PALS project were converted into guidelines for use in the Global Observations to Better the Environment (GLOBE) program so that GLOBE partners could align GLOBE curriculum activities to state science standards (<http://globeassessment.sri.com>). These

alignment guidelines could also be used by states to align assessments with state science standards and to create relational databases that would make access to standards-based items and tasks efficient.

**Develop Item/Task Banks.** The capacity of computers to retrieve large files with complex formatting permits assessment designers to create item/task banks. Since the late 1980s, such item/task banks have been created, but recently Web-based technologies enhanced their functionality. Item/task banks can be organized, classified, and stored according to the content and technical properties of the items and tasks. These banks are eclectic with regard to item formats and may include multiple-choice and open-ended items/tasks, as well as longer performance tasks. Oftentimes the items and tasks are classified by learning objective or standard. In those cases, a software program is used to identify items in the bank that are linked to particular objectives or standards.

Some states are using item bank software. Publishers have created an array of computer-based test products for low stakes assessments that incorporate banks of items, largely in the multiple-choice format. These item banks can be custom linked to a state's standards. Teachers search the item banks to assemble tests and quizzes customized to their curriculum and state standards. To date, these products tend to offer the easy-to-administer, easy-to-score multiple-choice questions and some constructed-response ones. Few of the products offer science items or items and tasks for challenging standards.

SRI International has developed a number of online assessment collections for science. Performance Assessment Links in Science (PALS), funded by the National Science Foundation (NSF) is a digital library of over 300 K-12 standards-based science investigations comprised of 2,500 items that are linked to the national science and mathematics standards, and to some state standards and science curriculum programs.

The GLOBE assessment web site was developed to offer classroom teachers templates and sample assessments for assessing students' ability to engage in integrated investigations using data collected in the Global Observations to Benefit the Environment (GLOBE) program.

The Council of Chief State School Officers (CCSSO) has developed a science item bank. The CCSSO State Collaborative on Assessment and Student Standards (SCASS) science collection is available on a CD-ROM. The collection offers over 1,500 items in the form of performance events, balanced test forms, and portfolio structures. (<http://www.ccsso.org/projects/SCASS/>). The recently funded AAAS item bank in science will develop items linked to strands in the *Atlas of Scientific Literacy* (Project 2061 & National Science Teachers Association, 2001). Sets of items will represent progressive sequences of conceptual development, along with prerequisite knowledge. [See <http://www.project2061.org/assessment.htm>].

The Principled Assessment Designs in Inquiry (PADI) project, as part of its online resources to support assessment design, is developing libraries of science inquiry tasks that will serve as exemplars of the assessment designs developed by the project, as well as items and tasks from other science inquiry projects. (See <http://www.padi.sri.com>)

Item/task banks have been developed by states, districts, and curriculum developers and publishers of commercial textbooks. Maine posts science assessments and scored student work from its state assessment system [<http://www.state.me.us/education/lsalt/LAS/>]. Other states have item banks, but they may not be accessible online. More often, sample tasks or released forms are intended to communicate information about the state testing content and structure.

States and districts have very elaborate assessment task specifications and creating or finding items to meet these specifications is very time consuming and costly. The use of technology-supported item/task banks permits states and districts to streamline the process of finding the items and, if the item bank is of sufficient size, create multiple forms of an assessment for administration. One scenario for how technology could support large-scale state assessment is as follows. A state could develop an item/task bank which may have several levels of access, including secure items/tasks and those that are released and accessible to teachers and/or students. Ideally, the online item/task bank would be supported on the Internet. State administrators could draw from the secure segments of the item bank to construct accountability measures. Teachers could draw from other segments of the item/task bank to create their own tests of key content (Quellmalz & Moody, 2004). The state of West Virginia (Willis, 1990) provided some early evidence about the use of the “item bank” concept in its financing of a bottom-up strategy to assess student outcomes. The state purchased one copy of testing software that allowed teachers to select items, construct their own tests, print them, and administer them to their students. The benefits of using the item/task bank included ease in generating tests for many uses and relieving teachers of the “busy work” of test construction. Recently the state of Georgia has offered two major testing programs via the Internet (Bennett, 2002). Teachers, district, and state administrators draw from these item/task banks to create tests, some of which guide classroom instruction and some of which are used for accountability measures.

These collections have conditions of use associated with them. Some give users the right to use the items or assessments as they are or to adapt the artifacts in the collection; other collections permit the user to examine the assessments and their items, but the user has to seek permission prior to use. Other collections only permit examination, but use cannot be granted.

Digital collections of assessments also differ widely in the kinds of supporting resources they provide along with the item collection. For example, PALS provides a relatively complete set of resources to support the use of its performance assessment tasks. PALS resources include: the item prompt; administration directions (including colorful illustrations of the equipment and materials needed for the administration of the tasks); scoring rubrics; samples of student work, and technical quality information when available. In addition, the PALS site contains guidelines for adaptation of the performance tasks and for conducting online rater scoring.

The NSF-funded Online Evaluation Resource Library (OERL) offers a digital collection of evaluation instruments (<http://oerl.sri.com>). Its accompanying resources include: the actual instrument; a specification of the topics covered by the instrument; item formats; numbers of items; contributor lists; professional development modules on how to develop instruments (i.e., surveys, interviews, observations, learning assessment) and how to design an evaluation. OERL also provides a glossary of terms and criteria for sound evaluation practice, including the selection, development, and use of evaluation instruments (Zalles, 2005; Fusco, Skolnik, Haertel et al. 2005). To promote awareness and use of assessments related to state standards and test

approaches, states are likely to develop more sophisticated dynamic online assessment collections.

### **Online Item Authoring and Revision Using Templates and Distributed**

**Collaboration.** Computers and Web-based technologies have many capabilities that can aid states and school districts in the efficient design and development of large-scale, standardized tests. The evolution and widespread availability of basic word processing, graphics, and spreadsheet programs facilitate state and district personnel, as well as teachers, in the development of their own items and sharing those items with others for the purpose of editing. Editing of items prompts, developing items formats, and selecting specific items from existing collections are assessment functions that are made much easier with general software and Web-based resources. Increasingly, software is available that is specifically devoted to the development of assessment items. These software programs contain item templates and may have ancillary resources associated with them such as libraries of mathematical and scientific notations, graphics, maps, and other supports that are not available with generic word processing tools. Such templates can support the generation of individual items, such as multiple-choice items with a closed format, more elaborate assessment tasks that involve a student completing several sequenced tasks, and the scoring rubrics that accompany the items and tasks. The GLOBE assessment templates illustrate such a resource. On the GLOBE assessment Web site, which is also cross linked to the PALS site, the template presents a form that can be used by the teacher to develop an assessment (<http://globeassessment.sri.com>) The template provides cues for inserting GLOBE data displays, an applied problem to which the data are relevant, and a sequences of stem questions related to the GLOBE assessment investigation framework. Another example of template-based item development is that used in the state of North Carolina (Bazemore, 2004).

An important technological advance in item authoring is the availability of Web-based editors that support real time editing of text. These editors are of interest to the assessment community because they can support the creative interplay between content and assessment specialists. Assessment specialists can identify flaws in item wording and examine the match between items and test objectives and standards. They can review the items taking into account the measurement model that will be required to score the assessment. Content specialists know the domain knowledge—facts, principles, concepts, and relationships among concepts, and skills--that need to be assessed in order to meet the assessment's objectives. This co-editing process eliminates a key concern that some assessment specialists have expressed in the past, when computer templates were first used to develop test items. Web-based resources permit co-design in real time with a team of diverse specialists who can be in geographically distributed locations. We no longer have an individual test developer filling out a template alone at her or his computer with out the benefit of timely and frequent feedback from others with complementary expertise. Once items have been created, they can also undergo revision using the same Web-based resources.

**Technical Quality Reviews by Remote Experts for Content/Bias.** During this activity, items in the pool for potential use are subjected to multiple reviews in order to guarantee that they meet selected criteria, such as accuracy of content coverage, grammatical correctness, and fairness for all students including those from diverse groups (i.e., sex, racial/ethnic identity, and

special needs). In addition, these items may be rated with regard to the alignment of items with relevant content expectations and frameworks, depth of knowledge assessed by the items, and alignment among each item's or task's constituent parts (i.e., item prompt, rubrics, and student work). Teachers at relevant grade levels may review items for their grade appropriateness, including their use of specialized vocabulary, reading load, use of graphics, technology load, manipulation of materials, and computation load. The results of such reviews can be used to refine the items prior to their actual implementation in the assessment system. Such reviews can be conducted online with raters entering digitized responses at distributed locations using shared workspaces.

**Assembly of Test Forms.** The large-scale assembly of test forms is easily supported by technology. Test forms are created by selecting items from a pool that are organized by a test blueprint or specification. The test blueprint specifies the number and types of items that are needed to measure each objective and a sampling strategy for selecting items. A relational database can be constructed that contains metadata for each item in the bank. A sophisticated selection strategy can be used to search the item pool and identify those items with the desired features. A technology-supported selection tool would apply rules that are associated with the “goals” and “targets” of the assessment being conducted. These rules would take into account variables required for scoring and constraints that relate to the assessment task design. (See Adaptive Testing for more detail on selection strategies to assemble test forms.) Thus, the technology-based tool will select one or more items from the pool of potential candidates and identify them for presentation and delivery.

Assessments can be assembled to test students at each level of a state's educational system (i.e., state-, district-, and classroom-levels). In such a case, it will be necessary to analyze the collection of items and tasks available at the different levels of the system, align them to standards, and select those that have the features necessary to elicit evidence of achievement on the standards. The technology-supported assembly tool will track which items have been used and the number of items needed to stock and replenish the item pools. Items will need to move from a secure to a released pool as they are used in repeated administrations of the assessments at different levels of the assessment system.

### **Design of an Articulated Assessment System**

In the following sections, we describe the kinds of technology that might be used to support large-scale assessment and provide examples of how technology can assist in the overall design of an articulated statewide assessment system (Quellmalz & Moody, 2004). We define an articulated assessment system as one which coordinates classroom and district assessments with a state-level assessment. A large, accessible item pool is central to this effort—a pool that can be used by classroom teachers and assessment designers at the local-, district-, and state-levels. Articulated assessment systems involve both formative and summative items in order to meet the range of assessment purposes that such a system encompasses. Formative items address the needs of classroom teachers who must make instructional decisions at several points along the way. Summative items are needed by the state-, district-, and school-level policymakers as a basis for accountability decisions.

**System Blueprint.** An assessment system blueprint that represents a multi-level, articulated system is a highly detailed resource that can be technology-supported. A statewide blueprint could represent the collection of items and tasks to be administered at the state, district, and classroom levels. The blueprint specifies the state content standards, the number of items available, items types, and a sampling strategy for selecting items at the indicator or objective level. If both summative and formative items are to be considered, then that classification must appear in the blueprint, as well. The content an item covers is specified by both its link to the content standard indicators/objectives and by the cognitive demands it places on the recipient.

The blueprint development is a cyclical activity where assessment designers consider both the length of the assessment and the particular reporting level (content, domain, indicator/objective) that is desired and the level of precision that is needed in student scores. If this is a multi-level assessment system, then consensus of the stakeholders in the assessment system would determine the level of detail to be documented in the blueprint. Assessment collection blueprints can be supported by relational databases which can be searched online.

**Technology-Based System Design.** The design of an assessment system requires both the collection and organization of domain-specific knowledge and skills and the specification of the assessment system's purpose, functions, and operational processes. Mislevy, Steinberg, and Almond (2003) describe a conceptual framework that supports the design of any assessment system, including a large-scale, articulated system. This framework includes identifying the "information, patterns, structures and relationships" that are required to organize assessment arguments and conceptualizing the assessment objects to be assessed (Mislevy, 2003). Domain analysis and modeling is what Mislevy and his colleagues call the processes by which the valued work, task features, representational forms, valued knowledge and skills, knowledge structures and relationships, and knowledge-task relationships are identified for the purpose of designing the assessment system. This information is then organized into a design structure (conceptual assessment framework) that is composed of student, evidence, task, assembly, and presentation models. Below are thumbnail sketches of three of the models described in Mislevy, Steinberg, and Almond that are essential to specifying an assessment argument and designing principled assessments:

*Student Model:* The knowledge, proficiencies, strategies, skills, and behaviors that will be measured by the assessment;

*Task Model:* The features of tasks that are needed to elicit evidence about the target proficiencies and the operations involved in authoring, calibrating, presenting and coordinating the assessment items and tasks;

*Evidence Model:* The part of the evidentiary argument that allows the test designer to reason from the observations in a task to a belief about what test takers know. The two parts to the evidence model are an evaluation component with rules for extracting evidence from the work products in terms of values of observable variables that have been identified; and a measurement component that specifies the statistical models that synthesize information from the observable variables over performances for values of the student model variables.

Mislevy and his colleagues describe assessment assembly and presentation models, as well, which have to do with the operational features of the assessment design system. All five of these models direct the operation of the assessment machinery—delivery of particular tasks, application of rubrics and evaluation rules, and the application of a measurement model, if desired. Mislevy argues that all design decisions have to be coordinated from the conception of an assessment system in order for the assessment to embody a complete and coherent argument that permits assessment designers to make the desired inferences about test takers' performances.

Currently, a technology-supported assessment design system based on Mislevy's work, including the creation of a student model, task model, and evidence model is being developed through the IERI-funded, Principled Assessment Designs in Inquiry (PADI) Project (Mislevy, Chudowsky, & Draney, et.al., 2003; Riconscente, Mislevy, Hamel & PADI Research Group, 2005). The online system incorporates the use of design patterns to lay out the assessment argument prior to beginning the formal design of the assessment system. A series of hierarchically-related templates are completed to support the specification of the student, task and evidence models. A demonstration software program entitled *Gradebook* (Hamel, Mislevy, & Kennedy, in press), illustrates how scores from tasks that were designed through the PADI design system can be accumulated and passed to an IRT-based scoring then to a calibration engine that estimates proficiencies specified in the student model. The IRT-based scoring engine works with the measurement model specified in the design system in order to represent complex assessment tasks and student models. The models the scoring engine support include multidimensional IRT models, item bundles to handle conditional dependence, and categorical-response models for ratings of complex performances. This scoring engine is based on the work of Wilson and his colleagues (e.g., Adams, Wilson & Wang, 1997).

To illustrate the role that online technologies play in this assessment design tool, we briefly describe the application that supports the design system. The PADI design system runs off a Web-based tool for modeling that creates and manipulates a simple data model without representing it in UML, but permitting collaboration and the use of examples to compare and validate the model. (Hamel & Schank, 2005). Referred to as Emo, this three-tier application permits domain experts without knowledge of UML, to discuss models of student understanding as they manipulate and compare several models in a distributed, shared workspace. Results of their modifications to the existing model of student understanding can be seen immediately through the use of a visual editor. As Hamel and Schank (2005) describe the application, the Web-based program exposes a "virtual" data layer, an abstraction between the model displayed and the real database structure (a node/relation implementation). Domain experts can use this exposed HTML layer to describe a set of assessment objects and relationship among the assessment objects. This online system is intended to make the process of assessment design more efficient, less costly, more educative for collaborators, and more likely to produce high quality assessments that have reproducible components that can be easily customized.

Technology can also support less sophisticated test design systems. Web-based test blueprints and task templates customized in statewide formats could be provided to guide development of assessment forms at district and classroom levels. Some states and publishers offer such forms to support construction of similarly structured tests for formative and summative purposes.

## Break-the-Mold Science Assessments

In this section, we describe innovative kinds of assessment tasks that could be designed using learning technologies to measure more complex forms of science content and inquiry processes. We propose assessment designs that derive from analyses of cutting-edge science curricula that are integrating the tools of science and information technology into K-12 instruction. Our analyses of these curricula sought ways to take advantage of the embedded technologies to structure formal assessment tasks and elicit evidence for the kinds of conceptual understanding and extended science inquiry promoted in the programs. In our view, many of the technologies embedded in science curricula can be used or adapted to elicit, collect, document, analyze, appraise and display kinds of student performances that have not been readily accessible through traditional testing methods. Furthermore, these technologies open the possibilities of ongoing, formative assessment of investigations-in-progress, in addition to the design of summative, accountability measures. We propose that a number of the technology applications that support science learning could be extracted, tuned, generalized, and re-purposed or re-designed for assessment purposes. We consider how the affordances of some of these technologies could be used to support more explicit, systematic student assessment.

By analyzing a sample of science curriculum programs that use technology to support sustained inquiry, we have identified tools embedded in these programs that could be exploited for assessment purposes as well. Technology applications have been used in a variety of curriculum projects to support different cognitive and metacognitive components of science inquiry. Figure 1 presents a conceptual model that relates these technologies to seven key components of project-based science inquiry curricula: (1) rich environments with authentic problems, (2) collaboration, (3) planning, (4) investigating, (5) analyzing and interpreting, (6) communicating and presenting, and (7) monitoring, evaluation, reflection, and extension. Technology applications have also been developed for assembling electronic notebooks or digital portfolios and digital libraries which can be used to: document an individual student's problem-solving efforts; archive data for a research team, classroom, or school; or collect examples of instructional resources, instructional activities, or assessments.

---

Insert Figure 1 about here

---

In the following section, we overview technology applications that have been implemented in a few lighthouse science inquiry projects. For each of the inquiry components in Figure 1, we consider how technology could be used for assessment purposes. Some of the assessments of inquiry described below (i.e., communicating/presenting; collaboration; monitoring, evaluation, reflection, and extension) may be more suitable for formative assessments embedded in science curricula and not as feasible for summative, large-scale assessments. However, as technologies support more interactive, dynamic assessment designs, the gap between classroom and large scale assessments may narrow.

**Rich Contexts and Environments.** Technologies are opening possibilities for students to engage in investigations that are beyond typical classroom resources. Students have access to

a wide range of physical phenomena through Web-based technologies, visualizations, microworlds, simulations, and microcomputer-based laboratories. Among science curricula that have employed technology to create powerful and engaging learning environments is Science Theater/Teatro de Ciencia. This project provided elementary students with a medium for exploring ideas about how things work, such as what makes rainbows, predator-prey relationships, or how tumors form (Lewis, 1996). In GenScope, technology enables students to explore the multi-level processes of genetics visually and dynamically, making explicit the causal connection and interactions between them (Horowitz, Schwartz, et. al, 1998). The Modeling Across the Curriculum (MAC) project engages students in simulations in physical and life science and in chemistry (Horwitz, Tinker, Dede, & Wilensky (2001). Authentic technology-based problem scenarios in these projects provide an inquiry context and scaffolding for extended investigations. For example, the Learning by Design project has developed overarching problem scenarios in physical science and earth science in which students design a vehicle that can navigate in several different kinds of terrain that might be found on the moon or construct a working model to show how they would manage to save a beach from destruction (Hmelo, Holton, & Kolodner, 2000). In the Journey to El Yunque project, upper elementary students can access video footage, data sets, and web-based content on the effects of hurricanes on the Puerto Rico rainforest (McGee, Corriess & Shia 2001).

One important advantage of technologically enhanced instructional materials is that these simulations, which employ interactive, visual formats, offer a promising second chance for students who have traditionally experienced difficulty grasping abstract scientific concepts. Yet, these rich, immersive environments tend not to be exploited for the design of systematic, curriculum-embedded student assessments or for standardized performance assessment tasks. Two notable exceptions are projects developed by the Technology Group at Vanderbilt (CGTV) and VideoDiscovery. The SMART assessments developed by CGTV periodically embed complex problems, or assessment challenges, within an ongoing investigation (Vye, et al., 1998). In addition, VideoDiscovery has developed science simulations of investigations typically inaccessible or inappropriate in classrooms such as conditions of plant growth or viral infections. These simulations are intended as student performance assessments and contain investigation tools that could be used by other development teams to design assessments for their inquiry curricula (Clark & Taylor, 1998).

In addition to providing the environment for authentic investigations and design problems, technology can support assessment activities by offering variants of problems and tasks which differ in complexity and levels of scaffolding. Thus technology can offer a range of environments tailored to the skill levels of students.

**Planning.** A number of curricula projects have developed digital templates to support student planning of investigations. These templates scaffold such activities as posing a problem, formulating a hypothesis, analyzing problem elements, and selecting investigation methods. An example is the Learning by Design project's Design Diary (Kolodner, Gray & Fasse, 2002). The SMART assessments developed by CGTV ask students to order investigation tools from catalogues, and then provide feedback on the appropriateness of their orders (CGTV, 2000). The WISE project scaffolds student's planning of investigations related to range of science issues (Linn & Clark, 2003).

As students engage in such planning activities, notebooks that organize and scaffold planning activities can save these plans in digital portfolios, then host on-line collaborations and conversations about the plans. Explicit assessments of students' planning strategies can take place through peer reviews and the development of criteria both as the plans are being developed, as well as during end-of-project reflections, upon the appropriateness and effectiveness of planning strategies.

**Conducting Investigations.** Visualizations, simulations, and measurement and data collection tools permit investigations typically inaccessible to classrooms. Technologies can make available to students the investigative tools used by scientists and provide access to processes that are not directly observable, are too expensive or dangerous, that take place too quickly or too slowly, or are on a scale that is too small or too large. For example, World Watcher and Global Observations to Benefit the Environment (GLOBE) are two projects that provide students with access to data sets and visualizations about weather and worldwide environmental systems. These curricula have transformed tools and techniques developed for scientists into environments to support students in the development of robust scientific understanding (Edelson, Brown, Gordin, & Griffin, 1999; the GLOBE Program <http://www.globe.gov>). Students actually analyze and interpret scientific data using a variety of technology-based tools. The visualizations used in these projects could also become components of performance assessments to test students' understandings and interpretations of the images as they relate to earth systems concepts. Such investigation tools can both support inquiry and produce records that allow assessments of students' appropriate use of the tools.

**Analysis and Interpretation.** A highly significant advance enabled by technology has been the development of analysis tools that allow interpretation of complex data sets and creation of descriptive and explanatory models. Spreadsheets and graphing tools are becoming widely available and could be readily incorporated into assessment activities. SimCalc offers a graphing calculator that other projects could use to support analyses and test if they have been done appropriately (Kaput, in press). Model-It is a tool embedded in multiple science inquiry curricula for constructing and testing models of complex, dynamic systems such as stream ecosystems (Melcalf, Krajcik, & Soloway, 2000). In a technology-rich secondary-level curriculum, the Modeling Across the Curriculum (MAC) project has developed interactive curricula for biology, chemistry and physics that pose problem-solving and exploratory challenges that can involve the manipulation of a computer model and the collection of data (Horwitz et al., 2001).

As students analyze data and evidence and formulate models, technology tools can support the analyses, in some cases rate their accuracy or appropriateness in comparison to expert models, and save the analyses for further discussion and appraisal. By offering usable, alternative modes for representing findings and conclusions, technology affords multiple modalities for students to show what they know and understand.

**Communication/Presentation.** Technology offers a variety of aids for formal publication and presentation of the results of investigations as well as communication during

investigations, as described in the section above on collaborative tools. General purpose and presentation programs such as PowerPoint can support communication and dissemination of scientific findings; however a number of scientific inquiry projects have developed customized publication tools to scaffold students' organization of ideas, evidence, explanations, representations of data, and conclusions. One example is SenseMaker which scaffolds students' organization of their questions, hypotheses, evidence and data, and conclusions for reports of investigations (Bell & Linn, 2000). For assessment purposes, these same tools can support formative conversations about the quality of learning visible in students' work or formal and on-line summative evaluations using established scoring rubrics.

**Monitoring, Reflection, Evaluation and Extensions.** In addition to supporting ongoing monitoring of investigations and design, some technology applications have been developed to support metacognitive reviews by students of their overall project inquiry or design strategies, outcomes, and possible extensions. Some of these projects have also structured technology-supported activities for the evaluation of reports and products. The Progress Portfolio, developed by the Supportive Inquiry-Based Learning Environment (SIBLE) Project is a general purpose tool that can be customized to structure the records and support conversations about students' reasoning during an investigation, data generated, and observations and conclusions (Loh, et. al, in press). The Design Diary prompts students to continually evaluate their designs according to their purpose, structure, function, durability, safety, ease of use, and development costs. In addition, JavaCAP, uses a case authoring tool to foster evaluation and reflection by organizing completed designs into four scenes combining text and images: Problem Presentation, Alternative Selection, Solution, and It's a Wrap (Kolodner, Gray, & Fasse, 2002). In ThinkerTools, the Inquiry Scorer, scaffolds judgments about the phases and structure of scientific inquiry represented in the students' project reports (Frederikson & White, 1998).

**Collaboration.** Collaboration tools can connect learners to vast resources, remote experts, and distant peers. Two discussion tools that have been developed are: "Web-SMILE" developed by The EduTech Institute's Learning by Design project; and "SpeakEasy" incorporated in the WISE program (Kolodner, et al., 2002; Linn & Clark, 2003).

Students' development of effective collaboration strategies can be assessed with and by these technologies. For example, as students conduct their research and investigations or initiate their designs, the technology could document the collaborations, automatically sample excerpts of interactions, and calculate frequencies and types of resources and experts accessed. These typically inaccessible data could form the bases of assessments of growth in students' collaboration strategies.

Thus, technology can organize and store students' iterative designs and solutions for ongoing monitoring and assessment, as well as for retrospective evaluation. As students and collaborators reflect on lessons learned, evaluate reports and designs and consider extensions, technology tools can support the development and use of scoring rubrics. Furthermore, the analysis and storage capacities of technologies can support interpretations of assessments and facilitate production of results for a variety of audiences.

## Conclusions

The assessment industry has entered an era that will be characterized by flexible forms that are technology-supported. To break the mold of on-demand, superficial testing, educational researchers and practitioners must join forces to exploit the affordances offered by available and emerging technologies and use these affordances to create assessment tasks for deep and expanded student science learning. Table 1 contrasts the procedures typical of traditional large-scale testing with the increased scope and flexibility of technology-supported assessment procedures. Assessment implementation involves replicable administration of items and tasks, collection and archiving of representative samples of performances, scoring and analyses of the quality of performances, and organization and reporting of results to myriad audiences. In the past, economics, logistics, and the paper/pencil format have largely limited testing procedures. In the 21<sup>st</sup> century, assessments can be administered to individuals or groups in diverse settings under conditions that can accommodate or scaffold the learners' entering levels. Student responses can be collected in digital formats for ready retrieval and analysis. Various forms of automated, categorical scoring will support qualitative and quantitative summaries of performance. Electronic portfolios and case libraries will collect and display student work and digitally link work to standards.

---

Insert Table 1 about here

---

Technology also greatly enhances the design possibilities for assessment tasks and items. We have described several technology-supported science assessments that were stimulated by analyses of learning technologies used in science curricula projects—but many of these same tasks and assessments can now be widely implemented and incorporated within large-scale assessments due to the affordances of technology. We summarize these assessment advances in the area of scientific inquiry in Table 2, which contrasts the item and task designs of traditional tests with the potential designs of technology-supported assessments. In the assessment of science inquiry, technologies can offer complex, rich problems that can be adapted to students' ability levels and prior knowledge. They can support collaborative research and problem solving and document students' strategies as they plan, design, and carry out research investigations. Technology can offer access to and present complex, vast content and resources for exploring research problems and conducting investigations. Simulations and a range of measurement tools vastly expand the kinds of science measurements and analyses students can perform virtually and in the field. Students can access, examine, and manipulate massive data sets presented in multiple representational formats such as visualizations and modeling tools. Spreadsheets and graphing tools can support analyses and displays of evidence and data. At the same time, technology can document these on-going strategies. Technologies can be used to organize and make multimedia presentations of findings from science investigations. Technologies can collect, organize, and scaffold students' metacognitive reflections on their progress as effective problem solvers, at the same time supporting systematic assessment of achievement through case libraries of student work appraised according to rubrics and accompanied by expert commentary.

---

Insert Table 2 about here

---

In this paper, we have described but a sample of the kinds of advances in technology that can improve student assessment in general and science assessment in particular. The technology supports for science assessments we urge both enrich and anticipate the growing technology infrastructures and capacities of today's classrooms. With these technologies, we believe our students can experience more fully the possibilities of education in the sciences in the 21st century.

States have embarked on the journey to electronic testing. They may be encountering bumpy roads as they attempt to build adequate technical infrastructures and secure catalytic funding. Some may view the shift of computer-based testing to low-stakes, district applications as a detour; others may see the burgeoning local market as an opportunity to stimulate instructionally relevant re-designs of current products. Testing in the domain of science offers both the opportunity and responsibility to design systems, tasks, and administration conditions that will elicit more valid measures of what students know and can do.

There is ample evidence that large-scale assessment is being transformed by technology. The greatest promise is to reinvent conventional assessment forms with technology supports and to offer policymakers, district administrators, students and teachers a world of new assessment genre. These forms, which seamlessly integrate all levels of the educational system, will advance current, large-scale assessment practices and guide instructional improvement at the program, school, and classroom level.

## References

- Adams, R. Wilson, M. R., & Wang, W.-C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*, 1-23.
- Almond, R. G., Steinberg, L.S., & Mislevy, R. (2002). A sample assessment using the four process framework. CSE Technical Report 543. Los Angeles: The National Center for Research on Evaluation, Standards, Student Testing (CRESST), Center for Studies in Education, UCLA.
- American Association for the Advancement of Science (AAAS). (1993). *Benchmarks for science literacy*. New York: Oxford University Press.
- American Educational Research Association, American Psychological Association, and the National Council of Measurement in Education (1999). *Standards for Educational and Psychological Testing* (2<sup>nd</sup> ed.). Washington, DC: Author.
- Baxter, G., Elder, A. D., & Glaser, R. (1996). Knowledge-based cognition and performance assessment in the science classroom. *Educational Psychologist, 31*(2), 133-140.
- Bazemore, M. (2004). Challenges with Technology-Based Assessments. Presented at the Secretary of Education's Accountability and Assessment Summit. St. Louis, MO.
- Bejar, I. I. (1996). *Generative response modeling: Leveraging the computer as a test delivery medium* (RR-96-13). Princeton, NJ: Educational Testing Service.
- Bell, P., & Linn, M. C. (2000). "Scientific arguments as learning artifacts: Designing for learning from the Web with KIE." *International Journal of Science Education, 22*: 797-817.
- Bennett, R. E. (1998). *Reinventing assessment: Speculations on the future of large-scale educational testing*. Princeton, NJ: Policy Information Center, Educational Testing Service.
- Bennett, R. E. (2001). How the Internet will help large-scale assessment reinvent itself. *Education Policy Analysis Archives, 9*(5), available online at [epaa.asu.edu/epaa/v9n5.html](http://epaa.asu.edu/epaa/v9n5.html)
- Bennett, R.E. (2002, Summer). Using electronic assessment to measure student performance. *The State Education Standard, 22-29*.
- Clark, D.J., & Taylor, S.N. (1998, April). *Multimedia simulations for science performance assessments*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

- Cognition and Technology Group at Vanderbilt (CGVT). (2000). Adventures in anchored instruction: Lessons learned from beyond the ivory tower. In R. Glaser (ed.), *Advances in Instructional Psychology* (Vol. 5, pp. 35-100). Mahwah, NJ: Erlbaum.
- Edelson, D. C., Brown, M., Gordin, D. N., & Griffin, D. A. (1999, February). Making visualization accessible to students. *GSA Today*, 9(2), 8-10.
- Education Week. (2003, May 8). Technology Counts 2003: Pencils down: Technology's answer to testing.
- Fredriksen, J. R. & White. B. Y. (1998). Teaching and learning generic modeling and reasoning skills. *Interactive Learning Environments*, 5, 32-51.
- Fusco, J., Skolnik, H., Haertel, G. D., Javitz, H., Smith, N., & Thurston, E. (2005). *OERL survey design, methodology, administration and results*. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.
- Hamel, L., Mislevy, R., & Kennedy, C. (in press). *A guide to the EDMS Gradebook*. (PADI Technical Report). Menlo Park, CA: SRI International.
- Hamel, L., & Schank, P. (2005). Participatory example-based data modeling in PADI. (PADI Technical Report No. 4). Menlo Park, CA: SRI International.
- Hmelo, C.E., Holton, D.L., Kolodner, J.L. (2000). Designing to Learn about Complex Systems. *Journal of the Learning Sciences*, Vol. 9, No. 3, pp. 47-298
- Horowitz, P., Schwartz, J. et al. (1998). *Implementation and evaluation of the Genscope™ learning environment: Issues, solutions, and results*. In Guzdail, M., Kolodner, J., & Bruckman, A. (Eds) (1998). *Proceedings of the Third International Conference of the Learning Sciences*. Charlottesville, VA: Association for the Advancement of Computer in Education.
- Horowitz, P., Tinker, R., Dede, C., Gobert, J., & Wilensky, U. (2001). *Modeling Across the Curriculum*. Interagency Education Research Initiative Grant funded by the National Science Foundation and the U.S. Department of Education.
- Kaput, J. (in press). Changing representational infrastructure changes most everything: The case of SimCalc, algebra and calculus. To appear in M. K. Heid & G. Blume (Eds.), *Research on Technology in the learning and teaching of mathematics: Syntheses and perspectives*. (with R. Schorr
- Kerrey, B., & Isakson, J. (2000). *The power of the Internet for learning: Moving from promise to practice- Report of the Web-based education commission*. Washington, DC: Web-based Education Commission.

- Klein, S. P., Hamilton, L., McCaffrey, D.F., Stecher, B. M., Robyn, A., & Burroughs, D. (2002). *Teaching practices and student achievement: Report of the first-year findings from the "Mosaic" study of systemic initiatives in mathematics and science*. Santa Monica, CA: RAND.
- Kolodner, J.L., Gray, J., & Fasse, B. B. (2002). Promoting Transfer through Case-Based Reasoning: Rituals and Practices in Learning by Design™ Classrooms. *Cognitive Science Quarterly*, Vol. 1
- Kreikemeier, P. A., Quellmalz, E., & Haydel, A. M. (2004). *Testing the alignment of items to the National Science Education Inquiry Standards*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.
- Lawrenz, F., & Huffman, D. (in press). The archipelago approach to mixed method evaluation. *American Journal of Evaluation*.
- Lewis, C. (1996). Science Theatre/Teatro de Ciencia. In project abstracts of the Applications of Advanced Technology Program Principal Investigator Meeting, June 27-28, Washington, DC.
- Linn, M. C., D. Clark, et al. (2003). "WISE design for knowledge integration." *Science Education*, 87: 517-538
- Loh, B., Reiser, B. J., Radinsky, J., Edelson, D. C., Gomez, L. M., Marshall, S. (in press). Developing Reflective Inquiry practices: A case study of software, the teacher, and students. In K. Crowley, C. Schunn, & T. Okada (Eds.), *Designing for Science: Implications from Everyday, Classroom, and Professional Settings*. Mahwah, NJ: Erlbaum.
- McGee, S., Corriss, B., & Shia, R. (2001). Using simulations to improve cognitive reasoning. Paper presented at the annual meeting of the American Educational Research Association, Seattle.
- Means, B., Penuel, B., & Quellmalz, E. (2002). Developing Assessments for Tomorrow's Classrooms. In W. Heinecke & L. Blasi (eds.), *Research Methods for Educational Technology. Volume One: Methods of Evaluating Educational Technology*. Greenwich, CT: Information Age Press.
- Melcalf, S.J., Krajcik, J., Soloway, E. (2000). Model-It: A design retrospective. In M.J. Jacobson & R.B. Kozma, (Eds.) *Innovations in Science and Mathematics Education: Advanced Designs for Technologies of Learning*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Minstrell, J. (1995). *Diagnoser: A Computerized Assessment System to Address Students' Understanding and Reasoning*. Paper presentation at the annual meeting of the AAPT, Spokane, WA.
- Mislevy, R. J. (2003). Substance and structure in assessment arguments. *Law, Probability & Risk*, 2, 237-258.

- Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1* (1), 3-62.
- Mislevy, R. J., Chudowsky, N., Draney, K., Fried, R., Gaffney, T., Haertel, G., et al. (2003). Design patterns for assessing science inquiry (PADI Technical Report No. 1). Menlo Park, CA: SRI International.
- Mislevy, R.M., Almond, R., Steinberg, L. Haertel, G.D., & Penuel, W. (2003). Leverage points for improving educational assessment. In B. Means & G. Haertel (Eds.), *Evaluating the effects of technology in education*. Hillsdale, NJ: Erlbaum.
- National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Reston, VA: author
- National Research Council (NRC). (1996). *National science education standards*. Washington, DC: National Academy Press.
- Neuberger, W. (2004). *Online Assessment in Oregon: The Technology-Enhanced Student Assessment (TESA)*. NCLB Leadership Summit. Saint Louis, MO. March, 2004.
- Odendahl, N. (1999). Online delivery and scoring of constructed-response assessments. Paper presented at the annual meeting of the American Educational Research Association, Montreal.
- Pellegrino, J., Chudowsky, N., & Glaser, R., (Eds.) (2001). *Knowing what students know: The science and design of educational assessment*. National Research Council's Committee on the Foundations of Assessment. Washington, DC: National Academy Press.
- President's Committee of Advisors on Science and Technology, Panel on Educational Technology. (PCAST, 1997). *Report to the President on the use of technology to strengthen K-12 education in the United States*. Washington, DC: Author.
- Project 2061 and National Science Teachers Association (2001). *Atlas of scientific literacy*. Washington, DC: American Association for the Advancement of Science, Project 2061.
- Quellmalz, E. S. (1984). Successful Large-Scale Writing Assessment Programs: Where Are We Now and Where Do We Go From Here? *Educational Measurement: Issues and Practices, 3*(1), 29-35.
- Quellmalz, E. S. (2003). *Performance Assessment Links in Science (PALS) Final Report*. Menlo Park, CA: SRI International.
- Quellmalz, E., & Haydel, A. M. (2002). *Using cognitive analysis to study the validities of science inquiry assessments*. Paper presented at the annual meeting of the American Educational Research Association Annual Meeting, New Orleans, LA.

Quellmalz, E.S. & Kozma, R. (2003). Designing assessments of learning with technology. *Assessment in Education*, 10 (3) 389-407.

Quellmalz, E. S., & Moody, M. (2004). Models for multi-level state science assessment systems. *Report commissioned by the National Research Council Committee on Test Design for K-12 Science Achievement*. Washington, DC: National Research Council.

Quellmalz, E.S. & Schank, P. (1998). *Performance Assessment Links in Science (PALS): On-Line Interactive Resources*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Quellmalz, E. S., & Zalles, D. (2002a). *Designing technology assessments: Cognitive-based modular design*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Quellmalz, E. S., & Zalles, D. (2002b). *World Links for Development: Student assessment Uganda field test, 2000*. Report to the World Links for Development Organization. Menlo Park, CA: SRI International.

Riconscente, M., Mislevy, R., Hamel, L., & PADI Research Group (2005). An introduction to PADI task templates. (PADI Technical Report No. 3). Menlo Park, CA: SRI International.

Russell, M., Goldberg, A., & O'Connor, K. (2003). Computer-based testing and validity: a look back into the future. In A. McFarlane (Ed.) *Assessment in education: principles, policy & practice, Vol 10, no. 3*.

Shermis, M.D. (2004). *Automated essay scoring*. Presented at the Secretary of Education's Accountability and Assessment Summit. Saint Louis, MO. March 2004.

Steinberg, L. S., Mislevy, R. J., Baird, A., Cahallan, C., Dibello, L. V., Centurk, D., Yan, D., Chernick, H. & Kindfeld, A. C, H, (2003). BioMASS. CSE Technical Report 609. Los Angeles: CRESST, University of California, Los Angeles.

Vye, N. J., Schwartz, D. L., Bransford, J. D., Barron, B. J., Zech, L., & The Cognition and Technology Group at Vanderbilt. (1998). SMART environments that support monitoring, reflection, and revision. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.) *Metacognition in educational theory and practice*. Hillsdale, NJ: Erlbaum.

Wayman, J.C. & Yakimowski, M. (2004, March). *Software to Facilitate Teacher Data Use and NCLB Reporting*. . Presented at the Secretary of Education's Accountability and Assessment Summit. Saint Louis, MO.

Webb, N. L. (1999). *Alignment of science and mathematics standards and assessments in four states. Research Monograph No. 18*. Madison, WI and Washington, DC: National Institute for Science Education and Council of Chief State School Officers.

Webb, N.L., Kaufman, J., Kaufman, D., & Yang, J. (2001). *Study of the impact of the statewide systemic initiatives program*. Technical report to the National Science Foundation. Madison, WI: Wisconsin Center for Education Research.

Whalen, S. J., & Bejar, I. I. (1998). Relational databases in assessment: An application to online scoring. *Journal of Educational Computing Research*, 18, 1-13.

Willis, J.A., (1990, summer). Learning Outcome Testing Program: Standardized classroom testing in West Virginia through item banking, test generation, and curricular management software. *Educational Measurement: Issues and Practice*, 9(2), 11-14.

Zalles, D. (2005). *Evaluating Web-based professional development*. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.

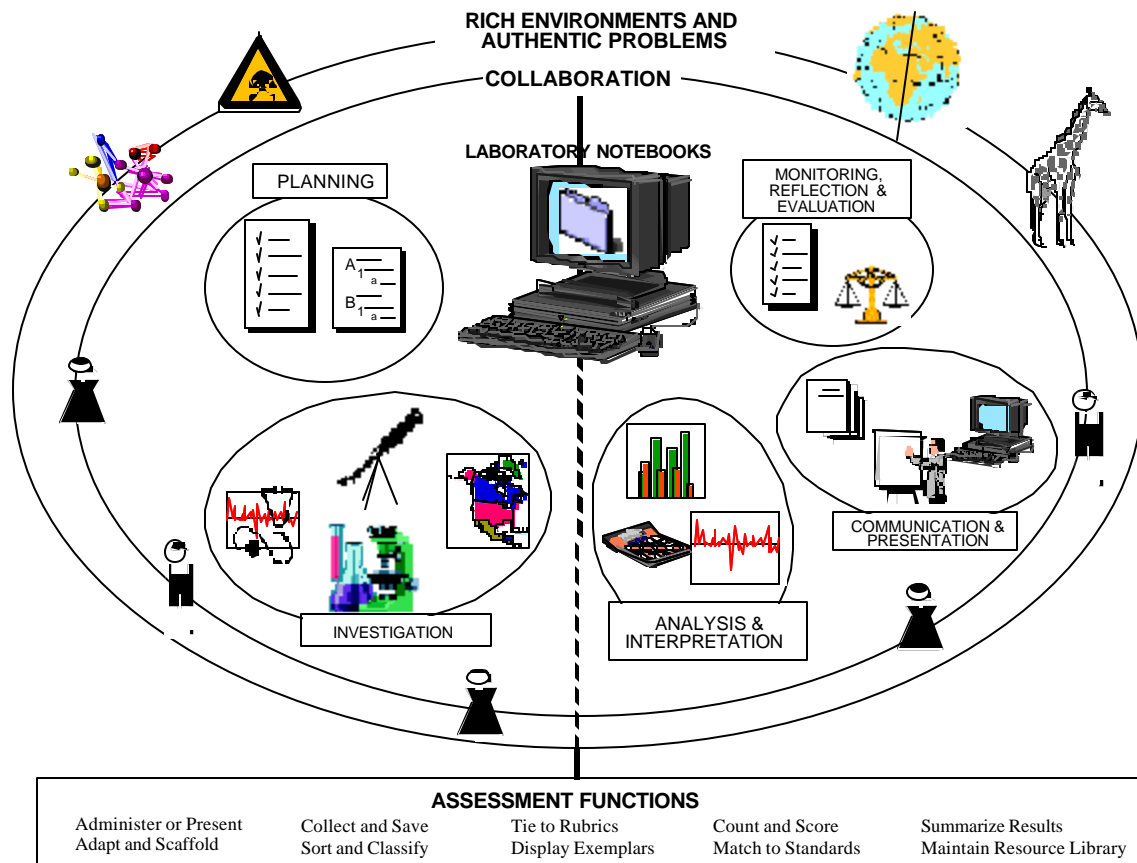


Figure 1  
Conceptual Model for Technology-Based Science Assessment

**Table 1**  
**Contrasts of Traditional Testing Procedures with Technology-Supported Assessments**

<b>Assessment Procedures</b>	<b>Traditional Assessments</b>	<b>Technology-Supported Assessments</b>
Administer	To individual learners  One common setting  Standardized conditions and procedures  Limited accommodations  On-demand, arbitrary timing  Annual  Summative	To individual learners or groups  Multiple, distributed settings, labs  Documented, flexible conditions and procedures  Extensive accommodations/scaffolding  Embedded, just-in-time  Ongoing or annual  Formative or summative
Collect and archive	Paper-pencil and optical scan formats  One time, one sample	Digital text archives  Digital video, audio archives  Internet-search traces  Collaboration records  Multiple collections, multiple samples  Digital portfolios
Score and analyze	Number correct or categorical ratings  Quantitative, cumulative data	Qualitative and quantitative scoring and interpretations  Automated scoring of natural-language responses (e.g., essays)  Coding/indexing of constructed responses, performances
Organize, record, and present	Graphical displays of scores and ratings  Some text work samples	Use of electronic portfolios to capture and score student progress recorded in text, audio, and video  Links to standards  Digital case libraries

**Table 2**  
**Contrast of Task/Item Designs in Traditional Tests with Technology-Supported Assessments of Scientific Inquiry**

<b>Scientific Inquiry Components</b>	<b>Traditional Testing Practice</b>	<b>Technology-Supported Assessment</b>
Contexts & problems	Decontextualized content  Discrete, brief problems	Rich, dynamic environments  Extensive Web-based resources  Access to remote experts  Extended, authentic problems  Scaffolded/adapted tasks and sequences
Collaboration	Typically prohibited	Directly assessed in ongoing documentation
Planning and design	Seldom tested, then with brief responses	Documented and appraised iteratively
Conducting investigations, collecting data	Infrequently tested  In performance tasks, limited to safe, economical, accessible equipment	Addressed in Web searches, hands-on tasks, simulations, and probeware
Analyzing and interpreting	Typically limited to small data sets and hand calculations	Possible to handle massive data sets, visualizations  Conduct complex multivariate analyses and display with spreadsheets, graphing tools
Communication and presenting	Occasionally, brief, written conclusions, reports	Support and documentation for ongoing informal communication, multimedia reports and presentations
Monitoring, evaluating, reflecting, extending	Typically not tested; if so, in brief written format	Documented and scaffolded by electronic notebooks, portfolios, on-line multimedia case libraries of student work rated according to rubrics with annotated commentary