

CHAPTER 3

21ST CENTURY DYNAMIC ASSESSMENT

**Edys S. Quellmalz, Michael J. Timms, Barbara C. Buckley,
Jodi Davenport, Mark Loveland, and Matt D. Silbergitt**

INTRODUCTION

How can we assess the core knowledge and skills that students need to succeed in the 21st century? Researchers and practitioners have identified a variety of competencies, referred to as 21st century skills, information communication technology (ICT) skills, media literacy, cyberlearning, and new literacies. With the explosion of information in all fields comes the need for students to become adept at using knowledge, not just memorizing it. Students must be able to apply complex skills such as problem solving, critical thinking, creativity, communication, and collaboration across a range of problems in academic and practical contexts. Thus, assessments of 21st century skills must provide students opportunities to demonstrate competencies for acquiring, applying, and transferring knowledge. We can tap into these skills through dynamic assessments that expand how phenomena, information, and data can be represented and increase the number of ways learners can demonstrate their knowledge and skills. In addition, technology can support use of scaffolding during assessment tasks to monitor and promote learning progress. The challenges for K–12 educators and

assessment developers are to create assessment tasks that allow students to demonstrate 21st century skills and to create evidence models for making inferences about student progress and proficiency.

In this chapter we focus on the design of dynamic assessments of cognitive learning in academic domains. We define 21st century dynamic assessment as assessments that capitalize on the affordances of technology to: (1) focus on complex, integrated knowledge structures and strategies; (2) provide rich, authentic task environments that represent significant, recurring problems; (3) offer interactive, immediate, customized, graduated scaffolding; and (4) analyze evidence of learning trajectories and proficiency. We synthesize research related to identifying significant 21st century target knowledge and skills, developing appropriate rich, interactive, assessment tasks, and eliciting evidence of development and achievement of 21st century skills that can inform teaching and benefit learning. We cite examples of assessments attempting to measure 21st century skills, then describe the SimScientists program as a case study for using technology-enhanced, dynamic assessments to measure complex science learning representing core 21st century skills. Specifically, we first describe how the evidence-centered design assessment framework has shaped the development of our assessments of 21st century skills. We address four key issues for educators wanting to assess these skills. First, what knowledge and skills do we want to assess? Second, what types of tasks and environments allow students to demonstrate proficiency in these skills? Third, what evidence models are required for making inferences about student progress and proficiency? We conclude with a discussion of the challenges we must address to realize the promises of dynamic assessment in the 21st century.

EVIDENCE-CENTERED DESIGN ASSESSMENT FRAMEWORK

The NRC report *Knowing What Students Know* presented advances in measurement science that integrate cognitive research findings into systematic test design frameworks (Pellegrino, Chudowsky, & Glaser 2001). Evidence-centered assessment design, described in depth in another chapter in this volume, is the process of creating assessments that meaningfully tap into specified targeted knowledge and skills. First, a *student model* specifies the domain knowledge and practices to be assessed. Then, a *task model* specifies the types of tasks or performances that will allow students to demonstrate the extent to which they can apply the competencies in the student model. Finally, the *evidence model* specifies the types of response summaries and scores that will indicate levels of proficiency (Messick, 1994; Mislevy, Steinberg, & Almond, 2003). Cognitively-principled assessment design for sci-

ence, therefore, would begin with a student model derived from a theoretical framework of the kinds of enduring science knowledge structures and strategies that are expected of students in a domain, provide problems and environments in which students can carry out tasks that demonstrate their proficiency, and make explicit what and how evidence from these tasks is being analyzed to gauge progressing proficiency.

Although the evidence-centered design framework may seem intuitive, in practice many simulation-based learning programs fail to identify the specific knowledge and skills that are targeted, and few assessments allow students the opportunity to demonstrate their understanding of rich interconnected knowledge and their ability to employ complex problem solving skills. Even when the targeted knowledge and skills are made explicit, the majority of assessment instruments remain static, traditional test formats that fail to provide an environment for students to demonstrate their ability to acquire, apply, and transfer knowledge in new settings. Finally, the evidence models (how the task performances are related to proficiency) are typically “point-based” and do not reflect partial, developing understandings or common misconceptions.

What 21st Century Knowledge and Skills Do We Want to Assess?

Schemas

Many of the targeted knowledge and skills specified in 21st century frameworks are not new to educators. Decades of research in cognition, measurement and psychometrics stress the importance of targeting integrated knowledge structures and models in assessments (Pellegrino et al., 2001). Across academic and practical domains, research on the development of expertise indicates that experts have acquired large, organized, interconnected knowledge structures, called schemas, and well-honed, domain-specific problem-solving strategies (Bransford, Brown, & Cocking, 2000). For example, expert writers use different schema when composing persuasive versus narrative texts. Mathematicians use schema to identify types of problems and initiate appropriate solution routines and strategies. Scientists use schema represented in physical, mathematical, and conceptual models as tools for generating and testing hypotheses and to communicate about natural and designed systems (Nersessian, 2008).

Ability to Acquire and Apply Knowledge

21st century skills imply that students not only have knowledge, but also have the skills to acquire, apply and transfer schematic knowledge in new contexts. What are the components of these complex skills? Research on

systems thinking and model-based learning suggest that effective learners form, use, evaluate, and revise their mental models of phenomena in a recursive process that results in more complete, accurate, and useful mental models (Gobert & Buckley, 2000). For example, students who participate in cycles of model-based reasoning build deeper conceptual understandings of core scientific principles and systems (Stewart, Cartier, & Passmore, 2005). Students with model-building proficiencies are able to interpret patterns in data and formulate general models to explain phenomena (Lehrer, Schauble, Strom, & Pligge, 2001). A growing body of research shows model-based reasoning to be a signature practice of the sciences, supporting how scientists create insights and understandings of nature through conceptual, physical, and computational modeling (Nersessian, 2008). Further, cognitive research shows that learners who internalize schemas of complex system organization—structure, functions, and emergent behaviors—can transfer this heuristic understanding across systems (e.g., Goldstone, 2006; Goldstone & Wilensky, 2008).

Thus, the framework for model-based learning provides a basis for identifying the domain knowledge and reasoning required for an integrated and extensible understanding of a system. Assessments can then be built around age appropriate simulation models of the components, interactions, and emergent behaviors characteristic of all complex systems, as well as the particular instances of these in the system being studied (Buckley, in preparation).

What Types of Tasks and Environments Allow Students to Demonstrate Proficiency in 21st Century Skills?

Affordances of Simulation-Based Environments as Learning Platforms

Technologies are seen as tools that support schema formation and mental model construction by automating and augmenting performance on cognitively complex tasks (Norman, 1993). In the domain of science, core knowledge structures are represented in models of the world built by scientists (Hestenes, Wells, & Swackhamer, 1992; Stewart & Golubitsky, 1992). Science simulations provide dynamic representations of spatial, temporal, and causal phenomena in science systems, often not directly observable, that learners can explore and manipulate. In contrast to animations, where students view predetermined scenes and can only control viewing direction and pace, simulations adapt the dynamic displays in response to learner inputs. Key features of simulations include manipulation of structures and patterns that otherwise might not be visible or even conceivable, representations of time, scale, and causality, and the potential for generating and superimposing multiple physical and symbolic representations. Moreover,

since simulations can illustrate content in multiple representational forms, simulations can inform students' mental models of concepts and principles and also reduce potentially confounding language demands. Simulations can present the opportunity for students to engage in the kinds of investigations that are familiar components of hands-on curricula as well as to explore problems iteratively and discover solutions that students might not have discovered in other modalities. Importantly, simulations also can make available realistic problem scenarios that are difficult or impossible to create in a typical classroom.

Numerous studies by multimedia researchers, cognitive psychologists, curriculum developers, and commercial companies illustrate the benefits of science simulations for student learning. Simulations can support knowledge integration and the development of deeper understanding of complex topics, such as genetics, environmental science, and physics (Hickey, Kindfield, Horwitz, & Christie, 2003; Horwitz, Gobert, Buckley, & Wilensky, 2007; Krajcik, Marx, Blumenfeld, Soloway, & Fishman, 2000; Doerr, 1996). For example, when Model-It was used in a large number of classrooms, positive learning outcomes based on pretest-posttest data were reported (Krajcik et al., 2000). After participating in the Connected Chemistry project, which used NetLogo to teach the concept of chemical equilibrium, students tended to rely more on conceptual approaches than on algorithmic approaches or rote facts during problem solving (Stieff & Wilensky, 2003). Seventh, eighth, and ninth grade students who completed the ThinkerTools curriculum performed better, on average, than high school students on basic physics problems and were able to apply their conceptual models for force and motion to solve realistic problems (White & Frederiksen, 1998). An implementation study of the use of *BioLogica* by students in eight high schools showed an increase in genetics content knowledge in specific areas, as well as an increase in genetics problem-solving skills (Buckley, Gerlits, Goldberg-Mansfield, & Swiniarski, 2004). Log files of student responses were analyzed to identify systematic (versus haphazard) inquiry performances, which correlated with overall learning gains (Buckley, Gobert, Horwitz, & O'Dwyer, 2010). At the middle school level, a simulation of an aquatic ecosystem was used to allow students to look beyond the surface structures and functions they could see when an aquarium served as a physical model. The simulation allowed students to create connections between the macro-level fish reproduction and the micro-level nitrification processes (Hmelo-Silver et al., 2008). Repenning and colleagues found that students using a collective simulation of multiple human systems significantly improved student learning about system facts, making connections, and applying knowledge about relationships between systems in new situations (Ioannidou et al., 2010). It appears that making the connections between system levels explicit ben-

efits students' understanding; dynamic simulations are a productive way to make those connections salient (Slotta & Chi, 2006).

Simulations and 3D immersive environments have been proposed from a "connectionist" view of mental processing to allow for the development of "projective identities" (Gee, 2008a), which can be utilized in a virtual environment to promote embodiment and reveal student thinking and problem solving abilities (Gee, 2008b). Embodiment through the use of immersive technologies can utilize multiple perspectives to trigger powerful psychological associations and facilitate transfer, resulting in a greater sense of empowerment and higher student engagement (Dede, 2009; Dunleavy, Dede, & Mitchell, 2009). Digital and networking technologies can connect learners with peers and experts, helping to build knowledge and to reveal its progression. Other 21st century skills that can be measured through the use of technology include argumentation (Squire & Jan, 2007), cooperation (Oberholzer-Gee, Waldfoegel, & White, 2010), collaboration (Hausmann, van de Sande, & VanLehn, 2008; Lim & Wang, 2005), and innovation (Gee, 2009).

Affordances of Simulation-Based Environments for Assessment

Given that simulations provide dynamic environments for learning in science, they can also support tasks that assess whether students are able to acquire, apply, and transfer knowledge through science investigations. Multimedia research suggests that when degrees of learner control and interactivity are variables, spatial representations allow students to create effective mental models and visualizations (Schwartz & Heiser, 2006). Rieber, Tzeng, and Tribble (2004) found that students who were given graphical feedback while using a simulation (laws of motion) that included short explanations far outperformed those given only textual information.

This interactive capability of simulations to provide contingent feedback and additional instruction in multiple modalities can be tapped in 21st century dynamic assessment task designs. Simulations are one resource for offering technology-enhanced ways to extend the 20th century methods of individually administered dynamic assessment tutorials. For example, Palincar, Brown and Campione (1991) developed a strategy of providing graduated prompts to scaffold completion of tasks. Their work built on Vygotsky's theory of the zone of proximal development between an individual's independent and guided performance (Vygotsky, 1987). Feuerstein and colleagues studied approaches for providing guidance during assessment (Feuerstein, Rand & Hoffman, 1979). Similarly, intelligent tutoring systems are providing graduated scaffolding in highly structured mathematics tasks, although primarily for instructional rather than assessment purposes (Feng & Heffernan, 2010).

In science, simulations can vastly expand feedback and coaching methods by offering and documenting types of graduated levels of graphical and textual feedback and coaching elicited by students' responses during simulation-based investigations. As in prior generations of dynamic assessment, types and levels of feedback and coaching are based on a student's response to assessment tasks carefully placed on a developmental continuum. 21st century dynamic assessments offer alternatives to labor intensive, one-on-one administration by using technology as the mechanism for administering the simulation-based assessments. The simulation program can link student responses to an underlying learning progression in order to offer appropriate levels of feedback and additional instruction designed to help move students along the task sequence. The assessment management infrastructure undergirding simulation-based assessments can be programmed to document the levels and amounts of help each student uses and to factor these into metrics for assessing and reporting learning progress. Moreover, such rich, dynamic, interactive assessment environments can vastly expand current conceptualizations of "adaptive testing." In sum, science simulations, in particular, open significant opportunities for the design of assessments of systems thinking, model based reasoning, and scientific inquiry as advocated in national science standards, but seldom tapped in static, conventional tests (Quellmalz, DeBarger, Haertel, & Kreikemeier, 2005).

Affordances of Technology for Universal Design and Accommodations

Technology-based assessments can have many of the flexible presentation and response features recommended in Universal Design for Learning (UDL) guidelines by reducing language demands that may interfere with understanding information and directions in assessment tasks and with expressing understanding. Assessments can make use of the flexibility provided by digital technologies to implement recommendations in the Universal Design for Learning framework: (1) representing information in multiple formats and media, (2) providing multiple pathways for students' action and expression, and (3) providing multiple ways to engage students' interest and motivation (CAST, 2008). The visual, dynamic, and interactive features of simulations can make assessment tasks more accessible to a greater range of students (Pellegrino et al., 2001). Computer technology makes it possible to develop assessments reflecting UDL and to embed these assessments in instruction. However, to avoid invalid estimates of student achievement, students should have prior experience using the technology-based assessments.

If these innovative formats are to be used for instruction and assessment, the needs of *all* students must be considered, particularly access for students with disabilities. Accommodations currently provided for print-based assessments have parallels in digital formats. The most commonly requested

accommodations are extended time, large print, and read-aloud. Extended time is important for a wide range of students, particularly those facing the additional cognitive demand of using the other accommodations. Large print has its parallel in digital assessments as a zoom feature that enlarges the screen. Text-to-speech (TTS) can replace read-aloud accommodations, useful for students with disabilities for whom print is a barrier to accessing the content of the assessments. TTS also has advantages for those English learners who may understand spoken English, but are not yet very proficient in written English (Kopriva, 2000). Thus, dynamic assessments can increase the opportunities for students to demonstrate proficiency in challenging science content and inquiry investigations for all students.

Affordances of Technology for Assessment Administration

Dynamic assessments allow for more flexibility in terms of when and where an assessment can be administered. Computer-based programs, such as SimScientists, offer on-demand access to assessments on a range of topics through classroom computers or a school's computer lab. Handheld technologies can increase the ability to deliver assessments in a wider range of spaces. Handhelds have been used in formal education to measure student learning in Jigsaw cooperative learning environments (Lai & Wu, 2006), inquiry-based science classrooms (Vonderwell, Sparrow, & Zachariah, 2005), outdoor learning spaces (Liu, Tan, & Chu, 2009), urban city centers (Morrison et al., 2009), class field trips (Weller, Bickar, & McGuinness, 2008), and historic and environmental sites (Klopfer & Squire, 2008). In addition to extending the temporal and geographic flexibility of administering assessments, handheld technologies also facilitate the collection of data and recording of student responses in these non-classroom environments (Patten, Sanchez, & Tangey, 2006).

Though formal education has been the focus of most technology-enhanced assessment development, digital technologies can also be used to measure learning in informal environments. Informal learning takes place in designed spaces (e.g., museums, science centers), natural environments, social settings, and in the home (NRC, 2009; NRC 2010). But how do we really know what people are learning in these environments? As paper and pencil are not appropriate in informal settings, many studies have examined ways that technology-based systems can assess how and what people are learning in informal spaces. Museums are developing electronic guide systems that deliver content and facilitate a museum visit, while also recording visitor responses in order to determine what they are learning (Bruce, 2010). These systems help museums to better understand how people learn from exhibits, help visitors connect to the museum space (Duff et al., 2009), and stimulate complex thinking about exhibit topics (Schmitt, Bach, Dubois, & Duranthon, 2010). Digital tools also enable education research-

ers to measure what people know about important topics that are not yet a part of established formal education curricula, such as nanotechnology (Crone, 2008) and climate change (Schultz & Shugart, 2007). In combination, the affordances of technology vastly expand what can be assessed and how, when, and where testing can occur.

What Evidence Models Are Appropriate for Assessing 21st Century Skills?

Increased Types of Evidence

Evidence models for the types of 21st century dynamic assessments envisioned in this chapter will be complex. There are several reasons for this. First, the number and range of constructs to be measured during assessment will increase. Future assessments will simultaneously be measuring a diverse set of variables such as the actions that students take in accomplishing the task at hand, numbers and types of coaching, responses to previous tasks, estimates of the difficulty of the tasks being undertaken, time taken to respond, and could even include biometric data such as emotional state (via video capture of expressions), galvanic skin response (via a wrist band) or posture (via a pressure pad on the chair) (Arroyo et al., 2009; D’Mello, Craig, & Graesser, 2009a, 2009b).

The actions and inputs that students make in response to complex tasks and items in the dynamic assessment process are gathered as raw data linked to different elements of each task. Raw data must be processed before being scored in the next stage of the assessment cycle. The response processing step identifies the essential features of the response that provide evidence about the student’s current knowledge, skills, and capabilities for identified content and inquiry targets.

For example, the Framework and Specifications for the 2014 NAEP Technology and Engineering Literacy assessments propose that when students are engaged in a design task, a student’s interaction with the tasks and tools in the scenarios will generate raw data. When processed, the student’s choice of tools to accomplish a design task, how the tools were used, and the design outcome all become quantifiable evidence of student proficiency. Other evidence will come from the selected-response or constructed-response questions that students answer during scenarios that include multiple, scenario-based tasks or in sets of shorter discrete items.

Evidence models for 21st century skills will need to accommodate not only a greater number of variables, but also the diverse nature of these variables. In traditional assessments, the data are processed into ordered categories such as correct/incorrect for selected responses or a score of 0, 1, 2, or 3 for constructed responses. Measures have followed a pattern in which

a higher ranking generally means that the student is doing better. But in dynamic assessments the measures are not necessarily ordered or linear. For some data in the evidence model, meeting a greater number of criteria might indicate higher achievement. However, greater values in the raw data will not always translate into higher scores. For example, a task may require students to test a system and make changes. Less knowledgeable students may test the system only once, failing to identify the optimal settings for the system. Slightly higher achieving students may use trial and error, testing the system many times in a non-systematic way. The highest achieving students may use a more strategic and therefore more efficient method. These students' responses would indicate higher achievement than students who had both fewer and greater numbers of tests. Thus, there need to be criteria for processing the response data in order to convert them to a form that is interpretable as part of the summary scoring.

The use of dynamic assessments for formative purposes presents another challenge for the evidence model. Formative assessments provide immediate feedback to students as they take the assessment, rather than waiting until a complete set of responses and actions are recorded. At the point that feedback needs to be given, we may have a very incomplete set of evidence on the current state of the student's knowledge and skills, and so traditional assessment methods that need evidence from many items to make reliable judgments have limited use. Methods for generating real-time feedback during dynamic assessments must rely on estimates of the state of student knowledge. In assessments meant to be used formatively, to intervene and adjust instruction, feedback can be accompanied by hints and coaching to provide additional instruction. The levels and amounts of coaching can become variables in analyzing developing proficiencies. This requires probability-based methods, which are discussed in the next section.

The desire to give immediate feedback to students will also push us to develop methods for interpreting written constructed responses, since we know from human-scoring of written responses that they provide deep insights into student thinking. Methods of Natural Language Processing are emerging that will enable the sort of interpretation and categorization needed to provide feedback to students in real time. There will be a growth of systems that can do this. Such technologies generally require that the system be "trained" by scoring hundreds of student responses with previously assigned scores (obtained by human scoring). Periodic checks by human scorers of subsequent computer scoring should also be instituted.

New Psychometrics Required for 21st Century Dynamic Assessments

Just as the diverse data types collected during dynamic assessments necessitate more complex evidence models, they also demand more sophisti-

cated psychometric analysis methods. Historically, the field of educational psychometrics has grown up around the types of responses that are typical in paper-based, large-scale assessments: primarily multiple-choice and written response items. Over the years, the methods of analyzing these types of student responses have become increasingly sophisticated, progressing from Classical Test Theory to Item Response Modeling methods that can model different dimensions of students' responses and even dynamically adapt the assessment to the ability of the student during the assessment.

However, the complex tasks envisioned for 21st century dynamic assessments cannot easily be modeled using just Classical Test Theory (CTT) and Item Response Theory (IRT). The complex tasks in dynamic assessments lead to diverse sequences of actions that produce multiple measures, often gathered simultaneously, and may involve interpreting patterns of behavior across multiple tasks. The multidimensional nature of assessment in simulations makes CTT unsuitable as a measurement method because it cannot model different dimensions of a performance simultaneously. When the assessments also need to be made in real time as the student is still engaged in the task, as is the case for classroom-based formative assessments, the interpretations are often calculated based on very limited amounts of data. The types of measurement methods that better lend themselves to simulations are probability-based methods (like IRT and Bayes Nets) that can handle uncertainty about the current state of the learner, can provide immediate feedback during tasks (e.g., Model Tracing or rule-based methods like decision trees), and are able to model patterns of student behavior (e.g., Artificial Neural Networks and Bayes Nets). These methods are briefly described below.

Item response models have the advantage that they place estimates of student ability and item difficulty on the same linear scale, measured in logits (a logarithm of the odds scale). This means that the difference between a student's ability estimate and the item difficulty can be used to interpret student performance. Since both the estimates of student abilities and the estimates of item difficulty are expressed in logits, they can be meaningfully compared. IRT could be a useful methodology to use in determining how much help students need in solving problems in an intelligent learning environment by measuring the gap between item difficulty and current learner ability (Timms, 2007).

Bayes nets have been widely used in intelligent tutoring systems, and over the years their use in systems for assessment has grown. For examples, see Martin and VanLehn (1995); Mislevy and Gitomer (1996); Conati, Gerner, and VanLehn (2002); Behrens, Frezzo, Mislevy, Kroopnick, and Wise (2008) and the example given later in this chapter.

Artificial neural networks (ANNs) have been widely used in intelligent systems, especially those in which the system needs to learn from data. An ANN is an adaptive, most often nonlinear system that learns to perform a function (an input/output map) from data. In science education, the work of Stevens in a series of projects in IMMEX (Interactive MultiMedia Exercises) provides an example of the use of ANNs. A recent article (Cooper & Stevens, 2008) describes the use of ANNs to assess student metacognition in problem-solving in chemistry.

Model tracing was developed for cognitive tutors produced by the Pittsburgh Advanced Cognitive Tutors (PACT) center at Carnegie Mellon University. Model tracing works by comparing the student's solution of a problem to an expert system for the domain of interest. Production rules, or rules about knowledge and skills in a given domain are, in this system, based on the work of cognitive scientist John Anderson. His ACT-R model represents skill-based knowledge (Anderson, 1993; Anderson & Lebiere, 1998).

Rule-based methods employ some logic method to decide how to interpret a student action in order to provide immediate feedback during formative assessments. A simple example would be posing a multiple-choice question in which the distractors (wrong answer choices) were derived from known misconceptions in the content being assessed. The student's incorrect response could then be diagnosed and immediate action, such as coaching, can be taken. This type of diagnosis is the basis of the work of Minstrell and Kraus in their work with the DIAGNOSER software that assesses students' knowledge in science and diagnoses their understandings and misconceptions (Minstrell & Kraus, 2007).

The increased use of dynamic assessments will need to be accompanied by psychometric methods suited to the complex nature of thinking and reasoning to be measured and the complex types of tasks in which 21st century skill must be applied. The designs and analyses of 21st century skill progressions will require collaboration of cognitive scientists, domain experts, and measurement experts.

HOW CAN DYNAMIC ASSESSMENTS OF 21ST CENTURY SKILLS BE INTEGRATED INTO BALANCED, MULTILEVEL ASSESSMENT SYSTEMS?

A critical issue arising from a focus on assessment of 21st century skills is the integration of dynamic assessments into assessment systems operating in states and districts to determine student proficiencies. The National Research Council (NRC) has developed a Framework for Science Education. States have set rigorous standards for what students should know and be

able to do. Yet, many states are still using traditional student outcome measures that are not tightly aligned with reform goals and do not document progress on challenging standards. State-level, large-scale tests typically favor breadth of content coverage over depth of reasoning. To address the need for stronger science testing methods, NSF funded the NRC project “Test Design for K–12 Science Assessment” to offer recommendations to states on their science assessment systems. In a report commissioned by that NRC project, Quellmalz and Moody (2004) proposed strategies for states to form collaboratives and use technology to create multilevel science assessment systems. They described how collaboratives can leverage resources, provide technology supports, and use assessment results at different levels of the system. Methodologies recommended in this report are applicable to non-science topics as well.

It is widely recognized that states must aim for balanced state assessment systems in which district, classroom and state tests are nested, mutually informative, and aligned (Pellegrino et al., 2001). Such assessment systems, like successful standards-based education, must be developmentally, horizontally, and vertically coherent (Wilson & Bertenthal, 2006). The *developmental coherence* of a system builds upon knowledge about how student understanding develops and the knowledge, abilities, and understandings that are needed for learning to proceed. The *horizontal coherence* of curriculum, instruction, and assessment arises from aligned learning goals and are mutually reinforcing. The *vertical coherence* of assessments at the classroom, school, district, and state can be forged with common goals for education as well as common purposes and methods of assessment.

The ideal model is based on alignment of assessments at each level with national and state standards based on a common set of design specifications for assessment tasks and items to be used at the various system levels (Quellmalz & Moody, 2004). Each level would employ assessment items and tasks that are common or parallel to those used at the other levels. Therefore, classroom assessments will use types of tasks and items parallel to those employed in district and state testing. In general, to serve formative, diagnostic purposes, classroom assessments would incorporate more items and complex assessment tasks than assessments developed for district and state summative, accountability purposes. The construction of assessments would vary for each level of the educational system, as would expectations for interpretation and use of the assessment results. However, the results would complement one another. While classroom assessments might provide immediate feedback to students and teachers about progress on a particular learning goal, statewide assessment would address proficiency on the learning goal, in the context of the larger set of related standards for all students.

Although less ideal, a more practical model may be needed where components of a system already exist such as those in multi-state collaboratives that already have unique state assessments and distinct design specifications. The newly funded Race to the Top assessment consortia have been funded to develop a range of assessment methods to test common core standards in English language arts and mathematics. In a balanced assessment model, vertical coherence is achieved through the use of reports that show the relationships among reporting categories at each level of the assessment system. Starting from a common set of standards, assessments at each level expand on the detail of inferences drawn at the level above. For example, at the state level, reports might describe achievement at the domain or sub-domain (strand) level. Within each strand, mastery of major topics might be reported at the district and school level. At the classroom level, progress toward mastery of these topics, along with immediate feedback to teachers and students, would still be possible.

Figure 3.1 shows how a report from multilevel assessments of science standards would connect related parts of the assessment system in either model.

MULTILEVEL ASSESSMENT REPORT

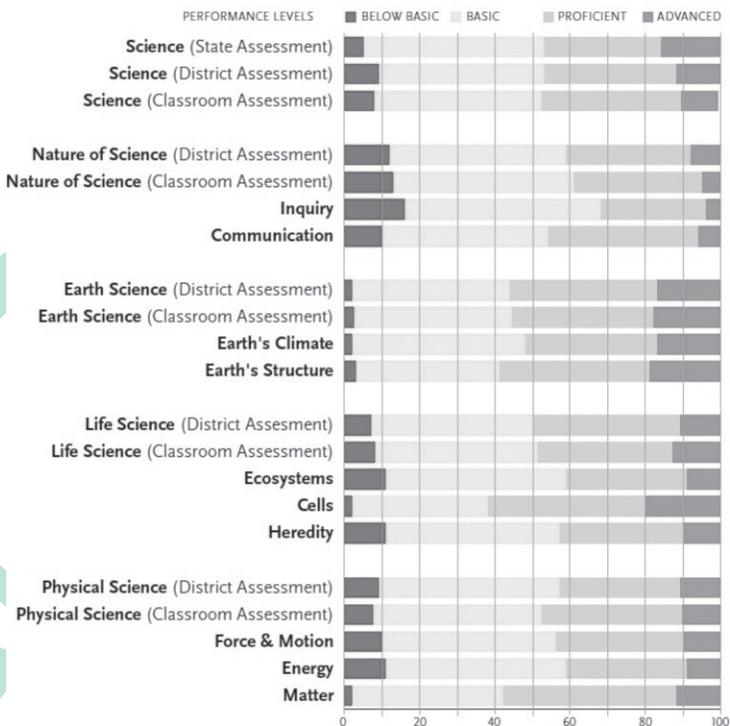


Figure 3.1 Report from a multilevel assessment system for science.

WHAT EXAMPLES EXIST OF ASSESSMENTS OF 21ST CENTURY SKILLS?

Assessments of 21st century skills require new open-ended, collaborative options for accessing, organizing, transforming, and communicating information. Programs, both small and large-scale, are beginning to explore the possibilities of dynamic, interactive tasks for obtaining evidence of learning achievement levels. Although the current accountability stakes and constraints tend to restrict a program's options, innovative designs are appearing in contemporary assessments. This new generation of assessments is moving beyond the use of technology for delivery and scoring of conventional item formats to harness technology that enables assessment of those aspects of cognition and performance that are complex and dynamic, and difficult or impossible to assess directly. Such work involves reconceptualizing assessment design and use, focusing in particular on relating, if not integrating, assessments more directly with learning environments. Some summative assessments are beginning to use interactive, scenario-based item sets, while curriculum-embedded assessments designed for formative purposes are beginning to use the affordances of technology to provide immediate, individualized feedback and graduated coaching during the assessment. The new generation of both summative and formative assessments will greatly expand the knowledge and processes targeted and the ways by which they are tested.

Summative Assessments

The majority of large-scale assessments are traditional paper and pencil tests, limited to multiple-choice format and open-ended items that require a written response. However, many knowledge domains and sub-domains are difficult to assess with traditional approaches. The potential for innovative assessment approaches is just now being considered. In English language arts assessments, for example, interactive, scenario-based tasks can address reading, writing, and discourse goals set within authentic problems. The 2011 NAEP for writing will employ computer-based prompts and word processing tools for students to compose a range of types of writing. Innovative dynamic assessments for English language arts could also offer students additional “tools of the trade” such as web search, highlighting, notepads, tables, and presentation tools to access, read, assemble, organize, transform, and represent information from multimedia resources composed of text, graphic, images, and video. Student responses could go beyond conventional item formats to include innovative computer-enabled formats

that employ hot spots, highlighting, cut and paste capabilities, table entry, written text, and presentation software.

For math performance assessments, interactive dynamic tasks could present multimedia images, graphics, and symbols along with technology-based mathematics “tools of the trade” to search and find data and information, analyze data, interpret or create visualizations, and use simulations to run iterative solutions, transform representations (tables, graphs), select and present best evidence, and present, explain, and display processes and solutions (Quellmalz, 2009).

Science assessment methods are leading the way, since knowledge of causal, temporal, and dynamic relationships among components within physical, life, and Earth systems, as well as inquiry processes, such as conducting investigations and communicating results, are difficult to test with traditional item formats (Quellmalz & Haertel, 2004). Some states have tested inquiry skills with hands-on performance assessments, but there are many logistical and economic challenges related to equipment, implementation, and scoring of such assessments both in classrooms and on the large scale required for state testing (Sausner, 2004). In their efforts to improve the validity and authenticity of assessments, many large-scale assessment programs are considering innovative formats made feasible by recent developments in computer-based testing. Both the Program for International Student Assessment (PISA) and the National Assessment of Educational Progress (NAEP) include interactive, computer-based components. For NAEP, interactive components are part of assessments for science, writing, technology and engineering literacy. In Minnesota’s state science tests, computer-based science assessments with innovative formats have been operational since 2008. In Utah, tryouts of Computer Innovative Items (CII) began in 2010. Along with increased opportunities to use technology in instruction and assessment, recent research and development of simulation-based assessments in science are providing evidence that simulations provide rich, real-world environments and test science knowledge and skills that tap the sorts of deep understanding that are difficult to test in paper format, or that are challenging to provide as hands-on tasks (Quellmalz, Timms, & Buckley, in press).

Formative Assessments

The benefits of formative assessment in the classroom are well established (Black & Wiliam, 1998, 2009; Nicol & Macfarlane-Dick, 2006); how technologies can support the necessary features of formative assessment is an ever-evolving area of research. From intelligent tutors to simulation-based curricula to video games and virtual worlds, information technolo-

gies are well suited to supporting many of the data collection, complex analysis, and individualized feedback and scaffolding features needed for the formative use of assessments. Considerable work has gone into the development of education technology for the assessment *of* learning, or summative assessment, as described previously. Assessment *for* learning has a different set of requirements, necessitating a slightly different approach to the development of technology-based assessment tools (Stiggins, 2006). Formative assessments should include an emphasis on authenticity and complexity in the content and methods of assessment rather than reproduction of knowledge and reductive measurement, as is typical of traditional classroom testing. Additionally, dynamic assessments are designed to serve formative purposes by facilitating meaningful feedback to students on “how they are doing,” providing additional scaffolding and instruction to bolster partial understandings, developing students’ abilities to direct their own learning and evaluate their own progress, and supporting the learning of others (McDowell et al., 2006). As a result, effective formative assessment can promote collaborative learning, dialogue and discourse, and the social construction of knowledge within a discipline (Sambell, 2010).

Technology-based assessments can provide extensive opportunities to engage in the kinds of tasks that develop and demonstrate student learning, building confidence and capabilities before students are summatively assessed. Using immediate, individualized feedback and customized follow-on instruction, dynamic assessments can provide coaching and hints when an error is detected (Quellmalz & Silbergliitt, 2010). When this feedback is graduated, students have multiple opportunities to confront their misconceptions with increasingly specific levels of coaching. The continuous use of this feedback has been found to help students revise their mental models of a given science system (Quellmalz & Silbergliitt, 2010). In addition to improving student mental models, technology-enhanced assessments for formative uses has also been shown to motivate and focus student learning, promote dialogical discourse, and develop metacognitive skills (Beatty & Gerace, 2009). Feedback can also come in the form of reports at the end of assessments. Reports that provide the kinds of descriptive feedback that help students connect their success in the assessment to their efforts are more productive than reports in the form of grades; the latter can undermine learning and student motivation (Covington, 1999; Maehr & Midgley, 1996).

The capacity to track student learning progress is particularly valuable for monitoring students with Individualized Education Plans. Computer technology also makes it possible to embed assessments reflecting Universal Design for Learning (UDL) in learning and assessment activities, and has been found to level the playing field for English language learners and students with disabilities (Wang, 2005; Case, Brooks, Wang, & Young, 2005;

Twing & Dolan, 2008). In the SimScientists assessments, this goes beyond the usual focus on text to include graphical representations, simulation controls, and an investigation of what happens when enlarging text or graphics results in loss of contiguity. Tools already built into students' computers can allow multiple representations and multiple media (Twing & Dolan, 2008; Case, 2008). Through the National Instructional Materials Accessibility Standard (NIMAS) and digital formats such as the Digital Accessible Information System (Daisy Consortium, 2006) automated transformation of text into alternate formats can be achieved (Twing & Dolan, 2008).

Below we describe how the SimScientists program uses simulations to push the frontiers of formative and summative assessment of science systems.

A Case in Point: The SimScientists Program

Funded by NSF, IES, and OSEA, projects in WestEd's SimScientists program (www.simsScientists.org) conduct research and development on the benefits of simulations for promoting and assessing complex science learning by developing powerful exemplars of formative and benchmark assessments of "new science literacies." The exemplars assess systems thinking, model-based reasoning, inquiry practices, and students' abilities to use the multimedia tools of science (Quellmalz, Timms, & Buckley, 2009; Quellmalz & Haertel, 2008). The SimScientists projects build on a theoretical framework that integrates model-based learning and evidence-centered design principles.

Models of Science Systems and Model-based Reasoning

Model-based learning (Gobert & Buckley, 2000; Buckley, in press) involves the formation, use, evaluation and revision of one's mental models of phenomena through a recursive process that results in more complete, accurate and useful mental models. Ideally, mental models build structures and relationships that represent science systems—the system structures (spatial arrangement of components), the interactions of those components, and the behaviors or properties that emerge from those interactions. Simulation-based assessments provide opportunities for students to demonstrate not only their understanding of a system, but also their ability to reason about the system as they predict, investigate, evaluate and explain the functioning of the system.

Thus, the framework for model-based learning provides a basis for identifying the domain knowledge and reasoning required for an integrated and extensible understanding of a scientific system. Instruction, investigations, and assessments can be built around simulations that represent the components, interactions, and emergent behaviors characteristic of all

complex systems, as well as the particular instances of these in the science system being studied.

SimScientists' assessment suites are composed of two or three curriculum embedded assessments that the teacher inserts into a classroom instructional sequence at key points for formative purposes and a summative benchmark assessment at the end of the unit. The dynamic embedded assessments provide immediate feedback and graduated coaching as students interact with the simulations, reports on progress to students and teachers, and off-line classroom collaborative reflection activities to help the teacher adjust instruction based on results of the formative simulation-based assessment. The summative benchmark assessment presents tasks and items parallel to those in the embedded assessments, but without feedback and coaching, to gauge student proficiency at the end of the unit. Students may work on the embedded assessments in pairs, but must take the benchmark assessment individually. Teachers are supported through face-to-face professional development, along with print and web-based guidelines, a procedures manual, help files, and the SimScientists HelpDesk. Assessments are delivered and data collected by the SimScientists' Assessment Management System (AMS).

SimScientists Student Model: Specification of System Models and Targets

SimScientists simulation-based assessments are being designed to address national middle school science standards related to science systems in life, physical and earth science. Simulation environments include ecosystems, biodiversity, human body systems, atoms and molecules, force and motion, climate, and plate tectonics. The assessments are aligned with national science frameworks and will be aligned with the new NRC national science frameworks and standards.

Using evidence-centered design methods, the design of each assessment suite begins with analyses of the domain, standards, and curricula. From these analyses, we define the three levels (components, interactions, emergent behaviors) that we will use to model the science system based on the grade-level-appropriate science standards. Figure 3.2 presents the system model specified for the middle school assessment of ecosystems.

The model levels in Figure 3.2 are represented in terms of food for energy and building blocks for growth and maintenance, organisms and their roles in dyad interactions (producers/consumers, predator/prey), as well as in the food web (diagrams that represent the flow of matter and energy through ecosystems). The population changes that emerge from interactions among organisms and with abiotic factors in the environment are represented in models that include both the organism view and graphs of populations. The last column in Figure 3.2 lists inquiry targets or science

Model Levels: Generic		Content Targets	Inquiry Targets
Components and Roles 	What are the components and behaviors of the system (at this level)? What are the “rules” of the system in general?	Every ecosystem has a similar pattern of organization with respect to the roles (producers, consumers, and decomposers) that organisms play in the movement of energy and matter through the system.	Use principles to identify role of organisms.
Interactions 	How do the interactions influence the individual components?	Matter and energy flow through the ecosystem as individual organisms interact with each other. Food web diagrams indicate the feeding relationships among organisms in an ecosystem.	Observe interactions among organisms.
Emergent Behaviors 	What is the overall state of the system that result from many interactions following specific rules?	Interactions between organisms and between organisms and the ecosystem’s nonliving features cause the populations of the different organisms to change over time.	predict observe explain investigate

Figure 3.2 SimScientists system model and content and inquiry targets—middle school ecosystems.

practices assessed at each level. The model levels described above—components, interactions, and emergent behavior—are ubiquitous in all systems.

SimScientists Task Models: Embedded and Benchmark Assessments

The cognitive demands of the student model are determined by the complexity of the assessment tasks and items. SimScientists task difficulties are affected by how phenomena are represented and the types of thinking and reasoning students must use. In SimScientists assessments, we base designs on progressively more complex models of science systems. We reframe inquiry standards in terms of the science practices and model-based reasoning needed for students to demonstrate and extend their understanding while conducting investigations at each level of a system. A progression of tasks both develops and elicits students’ understanding of the target models and inquiry skills. The tasks are designed to focus on inquiry and reasoning that resemble those of scientists as they create, observe, evaluate, and revise their models of phenomena. For example, we ask students to make observations to identify components and interactions, make predictions, design experiments, interpret data, evaluate their predictions, and explain the results and their reasoning. These are all key science practices. Cognition and multimedia learning research guide the design of student interactions with the simulations.

Curriculum-embedded formative assessment task models and evidence model. The dynamic assessment tasks designed to be used formatively must be sufficiently structured to enable targeted feedback and graduated coaching. Tasks that are too open make it difficult to provide useful feedback. Each is usually part of a series of steps necessary to complete a more complex task. As students complete each step, they receive one of four

levels of feedback and coaching. If correct, students receive confirmation and a restatement of the explanation. The first incorrect response triggers feedback to try again. The second incorrect response triggers feedback that restates the task and presents the rule or concept students need to apply to complete the task correctly. Feedback may also address common misconceptions. A third incorrect response triggers the correct explanation or action, and a worked example with detailed instructions for how to complete the task before students can move on. The examples shown in Figures 3.3 and 3.4 illustrate feedback and coaching provided in the dynamic embedded assessments for ecosystems.

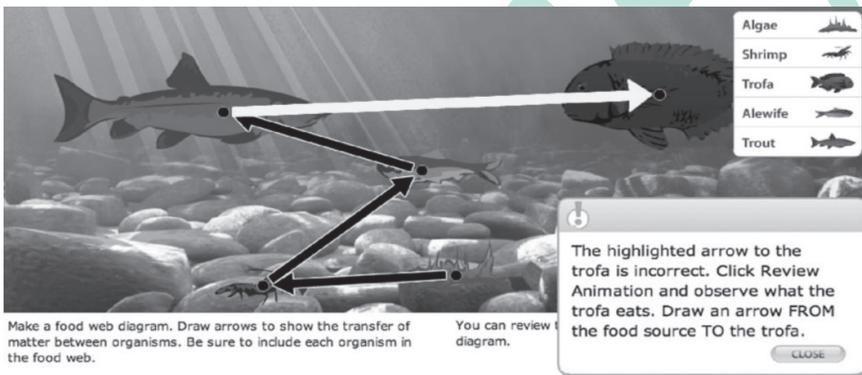


Figure 3.3 Screen shot of draw food web task with coaching.

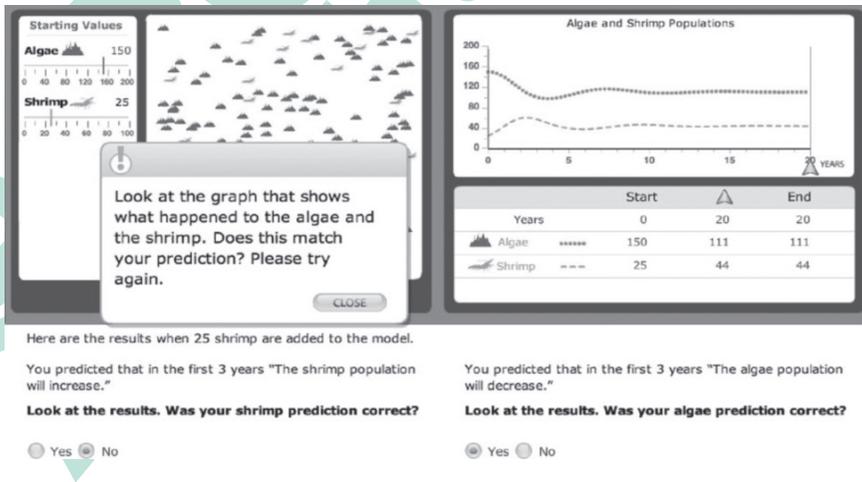


Figure 3.4 Screen shot of populations dynamics task with coaching.

In the task in Figure 3.3, students are asked to draw a food web showing the transfer of matter between organisms based on prior observations made of feeding behaviors in the novel ecosystem. When a student draws an incorrect arrow, a feedback box also coaches the student to observe the animation of feeding behaviors again and explains that the arrow should be drawn from the food source to each consumer.

Figure 3.4 shows feedback for a student asked to evaluate a prediction about emergent behaviors at the population level as depicted in the population graph. In this case, the student incorrectly evaluated his/her prediction and is coached to revisit whether the prediction and the data match and to try again.

When students have completed a dynamic embedded assessment, they receive a report on their progress (Figure 3.5). For both content and inquiry targets it describes the target knowledge or skill and the student's performance in terms of *on track*, *progressing*, or *needs help*. This is calculated based on the amount of help a student needed to complete the tasks related to that target.

After a class has completed a dynamic embedded assessment, the teacher accesses the Assessment Management System (AMS) to review student performances. The AMS provides a progress report for the whole class that summarizes this information and provides suggestions for grouping students into teams and groups based on their performances (Figure 3.6). Teachers are also encouraged to use their understanding of class dynamics to assign students to the suggested teams and groups.

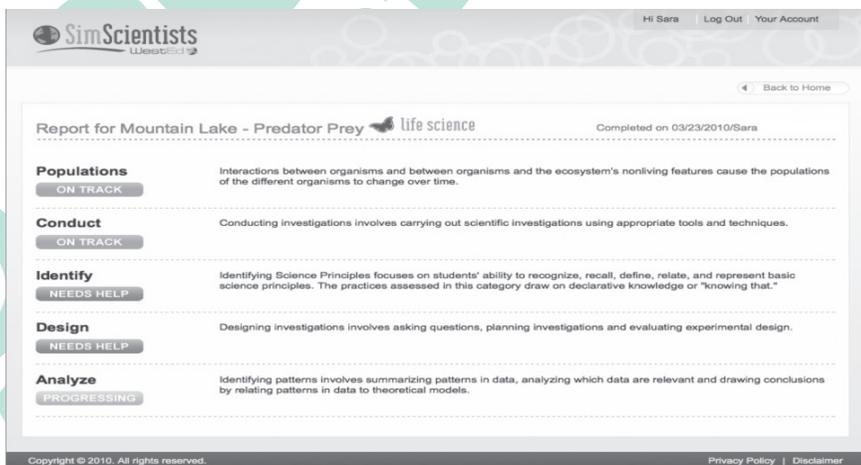


Figure 3.5 Student progress report for embedded assessment.

ASSESSMENT CLASS

Mountain Lake - Food Web Period 7 Go!

Needs Help Making Progress On Track

Reflection Activity PDF

Group A students needed little help on either roles or interactions
 Group B students needed help with interactions, but not with roles.
 Group C students needed help with understanding the roles of organisms in an ecosystem.

Student	Refl Gr	Roles	Interactions	Identifying	Using
Student 1	C	P	NH	NH	OT
Student 1	C	NH	NH	NH	NH
Student 3	A	OT	OT	OT	OT
Student 4	A	OT	OT	OT	OT
Student 5	C	NH	NH	NH	NH
Student 6	C	NH	NH	NH	P
Student 7	C	P	NH	NH	P
Student 8	C	NH	NH	NH	NH
Student 9	C	NH	P	NH	P
Student 10	B	OT	NH	OT	P

Figure 3.6 Class progress report for embedded assessment.

Embedded assessment reflection activities. An important component of the dynamic embedded assessment is an offline reflection activity designed to provide differentiated tasks and to engage students in scientific discourse as they apply their science content knowledge and inquiry skills to new, more complex ecosystems. Students are assigned to teams who are given tasks that address the content and inquiry targets with which they needed the most help. For example, one team might examine pictures of organisms eating behaviors to identify their roles as consumers. Another team might be responsible for identifying the producers. A third team might be responsible for drawing the arrows depicting the flow of energy and matter in the system. Small group work then feeds into a larger group task that requires a presentation to the class describing the roles of organisms in each ecosystem and the flow of matter and energy. Student peer assessment is promoted as the students as well as teachers evaluate the presentations using criteria for judging the evidence-base and clarity.

Summative, unit benchmark assessment task models and evidence model.

Tasks and items parallel to those in the embedded assessment are administered in the benchmark assessment. Benchmark tasks often combine component tasks of the embedded assessments into integrated tasks. Importantly, benchmark assessments require transfer of understanding of the model to a novel ecosystem (Figure 3.7) and do not provide feedback and coaching.

The SimScientists benchmark assessments employ a Bayes net to determine the proficiency levels of students on each of the content and inquiry targets. A Bayes net is a probabilistic graphical model that represents a set of random variables and their conditional independencies via a directed acyclic graph. In a Bayes net, nodes represent random variables and the edges (links between the nodes) encode the conditional dependencies be-

Grass
Cricket
Lizard
Kookaburra
Kangaroo

Make a food web diagram. Draw arrows to show the transfer of matter between organisms.

- To draw an arrow, click and drag from one dot to another dot.
- To delete an arrow, double click on it.

Be sure to include each organism in the food web.

You can review the animation and then return to this diagram.

REVIEW ANIMATION

Figure 3.7 Draw food web task in Benchmark assessment.

tween the variables. Across a series of nodes and edges a joint probability distribution can be specified over a set of discrete random variables. Figure 3.8 shows an example of a fragment of a Bayes Net used in the scoring of the ecosystems benchmark assessments in SimScientists. It shows how nodes in the network representing data gathered from student actions in the assessment (the lower two rows) provide information to assess the top-level variables of content knowledge and science inquiry skills represented in the upper two rows. Values for the edges are encoded, but not visible in this view.

Observable variables, that is responses that students gave and actions that they took in the simulation-based activities, are coded to the appropriate science content or science inquiry targets that they represent evidence of

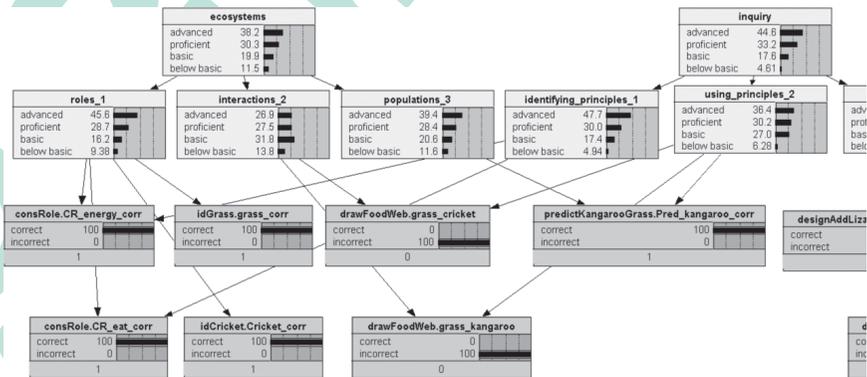


Figure 3.8 Fragment of a Bayes Net From the SimScientists Ecosystems Benchmark Assessment.

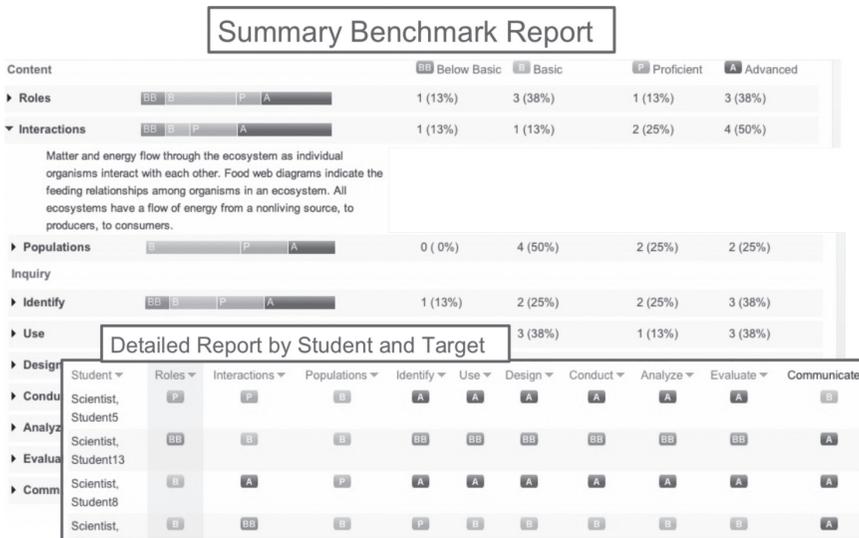


Figure 3.9 Benchmark reports for whole class and individual students.

in the student's performance. Using a scoring rubric provided in the learning management system (LMS), teachers score students' written responses. These scores are added to the record of observable variables for each student's assessment file. When all observable variables for an assessment are gathered, the teacher uses the learning management system to initiate scoring, a process that sends the observable variables data to the Bayes net. The data updates the probability estimates for each student in each of the content and inquiry targets and the report to the student shows the category (advanced, proficient, basic or below basic) that has the highest probability of being applicable for that student on each content and inquiry target. Results of the summative unit benchmark assessments are reported by the AMS in four proficiency levels for the content and inquiry targets.

Data on SimScientists Assessment Quality, Utility, and Feasibility

SimScientists projects have documented the effectiveness, utility, feasibility, and technical quality of simulation-based science assessments designed for curriculum-embedded formative assessment purposes and for summative accountability purposes. Calipers I, a demonstration project funded by NSF, documented evidence of the technical quality (validity and reliability), feasibility, and utility of simulation-based summative benchmark assessments for two middle school science topics (Quellmalz, Timms, & Buckley, in press). Alignments of the assessments with national science standards, as well as the accuracy of scientific content were con-

firmed by independent experts, including AAAS, and teacher reviews of the assessments. Teachers and students completed cognitive labs, thinking aloud as they responded to the simulation-based assessment tasks. The results provided construct validity evidence that the items elicited the intended science knowledge and skills. In-depth analysis of data from Calipers I demonstrated that simulation-based assessments can be developed to meet traditional psychometric measures for technical quality. An Item Response Theory (IRT) analysis of the response data from 109 students on the force and motion assessments and 81 students on the Ecosystems assessments showed that items functioned well overall. The mean reliability was .71 (Cronbach's Alpha), which for assessments that contain a mix of auto-scored, selected-response and human-scored constructed-response items, is within accepted usual ranges of reliability. The difficulty of the selected-response items ranged from .22 to .88, with a mean of .65, and the weighted mean square values of the items showed that all the items fit well to the science constructs being measured. In addition, by comparing the fit of one-dimensional and two-dimensional IRT models and testing if the difference was statistically significant using a Chi Square test with two degrees of freedom, it was shown that the assessment items in Calipers could be effectively used to measure two different dimensions of science content (content knowledge and inquiry skills), and that this approach yielded a more accurate measure of student ability than treating the science content knowledge as a single dimension. Evidence of the discriminant validity was based on the overall pattern of student performance that followed the expected progression of scores for high achievers, medium achievers, and lower achievers. Follow-up teacher interviews provided strong and enthusiastic support for the utility and feasibility of the simulation-based assessments.

Similar research to that conducted in Calipers I is currently underway in the NSF-funded Calipers II project on the benefits of simulation-based *formative* assessments that are embedded in curriculum units. The technical quality of linked unit benchmark assessments is also being documented. Data collection on the feasibility, utility, reliability, and validity of the formative and summative assessment modules continues in Calipers II. To date, data have been collected for two science topics, including expert reviews by AAAS of scientific accuracy, alignments of the assessments with national science standards and representative curricula, cognitive laboratories with teachers and students, classroom pilot testing, classroom observations, and teacher surveys and interviews. Data from 28 think-aloud sessions have provided preliminary evidence of the construct validity of the assessment tasks for eliciting the intended knowledge and skills. Classroom observations de-

tected some initial technical deployment challenges that were overcome. Observations also showed high student engagement. The embedded assessments took students about 15–30 minutes and the benchmark assessments took about 20–40 minutes to complete.

In a third project, *Integrating Simulation-Based Science Assessments into Balanced State Science Assessment Systems, an Enhanced Assessment Grant (EAG)*, funded by the Office of Elementary and Secondary Education, six states are involved in a study in which we developed three types of computer-based accommodations (visual enlargement, text-to-speech and extended time) which were added to the Calipers II simulation-based formative and benchmark assessments for two topics: ecosystems and force and motion. We have pilot tested them in approximately 60 middle school classrooms with 6,000 students to determine if these richer, deeper forms of science assessment benefit student learning, profile student proficiencies in more depth, and can augment evidence of achievement of standards in a state science assessment system (Quellmalz & Moody, 2004). The resulting data set will allow item response model analyses of the performance of the assessment tasks and items to further establish the technical quality of the simulation-based assessments and examine how well they measure performances of students with disabilities and of EL students who use the accommodations. As of May 1, 2010, the assessments for ecosystems and force and motion have been implemented in three states by 55 teachers with 5,867 students in 39 schools in 28 districts. Two hundred eighteen students (5%) have used the accommodations (mostly the text-to-speech and extended time). Responses from the teachers on online teacher surveys indicated that, overall, teachers were able to use the assessments successfully in their instruction and that, with occasional assistance from SimScientists help desk, were able to overcome technology challenges and complete the assessments. Teachers rated the quality and utility of the assessments highly. In addition, UCLA CRESST, the external evaluator, conducted nine case studies to evaluate use of the assessments in more depth. Data will be analyzed during the summer of 2010. Six states (NV, UT, NC, CT, MA, and VT) are participating on the project Design Panel to examine the potential of including the summative simulation-based unit benchmark science assessments in a state science assessment system. Utah is currently pilot testing short science simulation tasks for the online state science test. Data from this EAG project is providing strong evidence that the simulation-based embedded and benchmark assessments can be used on a large scale with diverse student populations and in school systems with varying technical infrastructures.

WHAT ARE THE CHALLENGES AND PROMISE FOR THE FUTURE?

Challenges

Dynamic assessments of 21st century skills are on the increase, but still relatively rare. One problem is definitional. Frameworks for 21st century skills range across disciplinary, workplace, and societal domains. Debates continue about the relative value, practicality, and benefits of assessing generic skills versus domain-specific knowledge structures and strategies, individuals versus groups, technology tool operations, and methods for appraising less tractable constructs such as innovation, creativity, citizenship, and responsibility. Moreover, new and updated frameworks and standards present moving targets for assessment design. In this chapter, we have focused on 21st century cognitive constructs such as problem solving, communication, and collaboration as they apply in academic disciplines, particularly in model-based learning and reasoning in science.

Although rhetoric often calls for challenging, authentic complex tasks within which learners apply 21st century skills, the range of contexts and tasks in current assessment practice is quite limited. The high stakes typically associated with summative purposes tend to suppress innovation in favor of brief, highly structured assessment tasks targeting simple concepts and skills and collecting standard selected and constructed responses. Concomitantly, assessments intended for formative classroom uses tend to mirror the limited constructs and task designs in summative assessments. Too often, engaging, rich technology-based learning environments do not explicitly articulate learning goals, present dynamic embedded assessments to promote learning, or use the affordances of the technologies to gather evidence of attainment of those goals. Consequently, there are relatively few exemplars, some described in this volume, of rich, dynamic tasks designed to assess significant, specified 21st century skills.

At the same time, dynamic assessments produce extensive log files of learner responses. The assessment framework of evidence-centered design calls for specification at the outset of the data that will provide evidence of achievement of targeted knowledge and skills. However, many technology-based learning and assessment environments defer principled extraction of evidence of learning in favor of exploratory post hoc mining of log files. In this chapter, we assert that dynamic assessments must develop new evidence models and psychometric methods to extract data from complex assessment tasks that can be combined in various ways to characterize and monitor the progressive development of 21st century proficiencies.

Implementations of these novel dynamic assessments pose a number of challenges. Teachers and students are likely to require training and expe-

rience to take advantage of the new dynamic assessments. Students need to become familiar with the novel formats and their operations. Teachers need professional development on identifying appropriate places to use the assessments formatively within instruction, and on methods for implementing and interpreting them. Policymakers and the public need information on the purposes, features, and benefits of dynamic assessments.

A final challenge to scalability and sustainability of dynamic assessments is the cost of initial development, ongoing maintenance, updates, and changes in technology requirements and platforms. Considerable development costs of dynamic assessments may be defrayed by designs that allow the assessments to test standards addressed in multiple curricula and reusability of designs and components. Research must provide evidence that the benefits of dynamic assessments warrant the costs.

Promise

Dynamic assessments of 21st century skills promise to revolutionize the types of learning that can be assessed and the ways in which learning can be measured, promoted, and interpreted. Broad 21st century cognitive capabilities such as problem solving, inquiry, communication, collaboration, and tool use can be tested within rich, authentic, complex contexts and tasks. Learners can employ significant, extended strategic thinking and reasoning, using a range of technology “tools of the trade.” Rich assessment tasks can present highly engaging 2D simulations and 3D virtual worlds in which collaboration and discourse play key roles in developing solutions and achieving goals. Assessment tasks can present scenarios that will both test and promote, through feedback and graduated coaching, the development of schema and mental models that learners can transfer across prototypical problems in academic and applied domains. Systems thinking and model-based reasoning can become manageable assessment targets. The dynamic nature of the new generation of assessments will further open the types of phenomena and environments that can be presented in assessment problems and the opportunities to document evidence of interest and engagement.

The distinctions between learning and assessment will become blurred as dynamic systems can provide immediate feedback and offer customized coaching and learning opportunities. Conceptions of adaptive testing can move from simple branching based on item difficulty statistics to multiple learning progressions based on cognitive analysis of the development of domain knowledge and skills. As we start to collect more and more evidence of learning and incorporate dynamic feedback and coaching into the assessments, they will become more like intelligent tutoring systems that can gauge and scaffold performance in rich, complex tasks. As these trends

progress, we will see more blending of the methods from both the fields of educational measurement and of intelligent tutoring to form a hybrid system in which learning and assessment are blended in such a way that they are indistinguishable.

Dynamic assessments can be delivered by a variety of platforms, allowing for more flexibility in terms of when and where evidence of learning can be collected. Dynamic assessments can be administered on school-based computers and also by ever-changing mobile computing devices. Portability will enable assessments in informal environments in designed educational spaces such as museums or in distributed out-of-school learning communities. Dynamic environments are well on their way to fulfilling the promise to transform assessment of learning.

REFERENCES

- Anderson, J. R. (1993). *The rules of the mind*. Hillsdale, NJ: Erlbaum.
- Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Arroyo, I., Woolf, B. P., Cooper, D. G., Burleson, W., Muldner, K., & Christopherson, R. (2009). *Emotion sensors go to school*. Paper presented at the International Conference of Artificial Intelligence in Education, Brighton, England.
- Beatty, I., & Gerace, W. (2009). Technology-enhanced formative assessment: A research-based pedagogy for teaching science with classroom response technology. *Journal of Science Education and Technology*, 18(2), 146–162.
- Behrens, J. T., Frezzo, D., Mislevy, R., Kroopnick, M., & Wise, D. (2008). Structural, functional, and semiotic symmetries in simulation-based games and assessments. In E. Baker, J. Dickieson, W. Wufleck & H. F. O'Neil (Eds.), *Assessment of problem solving using simulations* (pp. 59–80). New York: Lawrence Erlbaum Associates.
- Black, P., & Wiliam, D. (1998). *Inside the black box: Raising standards through classroom assessment*. London, UK: King's College.
- Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability*, 21(1), 5–31.
- Bransford, J. D., Brown, A. L., & Cocking, R. R. (2000). *How people learn: Brain, mind, experience, and school*. Washington, DC: National Academy Press.
- Bruce, J. (2010). Innovative museum guide systems. Paper presented at the Designing Usable Systems: 2010.
- Buckley, B. C. (in preparation-a). Model-based learning. In N. Seel (Ed.), *Encyclopedia of the sciences of learning*. New York: Springer Science.
- Buckley, B. C. (in preparation-b). Supporting and assessing complex biology learning with computer-based simulations and representations. In D. Treagust & C.-Y. Tsui (Eds.), *Multiple representations in biological education*. New York: Springer Science.

- Buckley, B. C., Gerlits, B., Goldberg-Mansfield, A., & Swiniarski, M. J. (2004). *The impact of biological usage in classrooms on student learning outcomes*. Paper presented at the National Association for Research on Science Teaching, Vancouver, BC.
- Buckley, B. C., Gobert, J., Horwitz, P., & O'Dwyer, L. (2010). Looking inside the black box: Assessing model-based learning and inquiry in Biologica. *International Journal of Learning Technologies*, 5(2).
- Case, B. J. (2008). Accommodations to improve instruction, learning, and assessment. In R. Johnson, & M. Ross (Eds.), *Testing deaf students in an age of accountability*. Washington, DC: Gallaudet Research Institute.
- Case, B. J., Brooks, T., Wang, S., & Young, M. (2005). *Administration mode comparability study*. San Antonio, TX: Harcourt Assessment, Inc.
- CAST. (2008). *Universal design for learning guidelines version 1.0*. Wakefield, MA: Author.
- Conati, C., Gernter, A., & VanLehn, K. (2002). Using bayesian networks to manage uncertainty in student modeling. *Journal of User Modeling and User-Adapted Interaction*, 12, 371–417.
- Cooper, M. M., & Stevens, R. (2008). Reliable multi-method assessment of meta-cognition use in chemistry problem solving. *Chemistry Education Research and Practice*, 9, 18–24.
- Covington, M. V. (1999). Caring about learning: The nature and nurturing of subject-matter appreciation. *Educational Psychologist*, 34(2), 127–136.
- Crone, W. (2008). *Bringing nano to the public through informal science education*. Paper presented at the 2008 American Physical Society March meeting.
- D'Mello, S. K., Craig, S. D., & Graesser, A. C. (2009a). Automatic detection of learner's affect from gross body language. *Applied Artificial Intelligence*, 23, 123–150.
- D'Mello, S. K., Craig, S. D., & Graesser, A. C. (2009b). Multi-method assessment of affective experience and expression during deep learning. *International Journal of Learning Technology*, 4, 165–187.
- Daisy Consortium. (2006). Digital accessible information system (DAISY). Available from <http://www.daisy.org>.
- Dede, C. (2009). Immersive interfaces for engagement and learning. *Science*, 323(5910), 66–69.
- Doerr, H. M. (1996). Integrating the study of trigonometry, vectors, and force through modeling. *School Science & Mathematics*, 96(8), 407.
- Duff, W., Carter, J., Dallas, C., Howarth, L., Ross, S., Sheffield, R., & Tilson, C. (2009). *The changing museum environment in North America and the impact of technology on museum work*. Paper presented at Empowering users: An active role for user communities.
- Dunleavy, M., Dede, C., & Mitchell, R. (2009). Affordances and limitations of immersive participatory augmented reality simulations for teaching and learning. *Journal of Science Education and Technology*, 18(1), 7–22.
- Feng, M., & Heffernan, N. (2010). Can We Get Better Assessment From A Tutoring System Compared to Traditional Paper Testing? Can We Have Our Cake (Better Assessment) and Eat It too (Student Learning During the Test)? Educational Data Mining, 2010.
- Gee, J. P. (2008a). What's a screen mean in a video game? Paper presented at the 7th international conference on Interaction design and children.

- Gee, J. P. (2008b). Video games and embodiment. *Games and Culture*, 3(3–4), 253–263.
- Gee, J. P. (2009). Pedagogy, education, and 21st century survival skills. *Journal of Virtual Worlds Research*, 2(1), 4–9.
- Gobert, J. D., & Buckley, B. C. (2000). Introduction to model-based teaching and learning in science education. *International Journal of Science Education*, 22(9), 891–894.
- Goldstone, R. L. (2006). The complex systems see-change in education. *The Journal of Learning Sciences*, 15, 35–43.
- Goldstone, R. L., & Wilensky, U. (2008). Promoting transfer through complex systems principles. *Journal of the Learning Sciences*, 17, 465–516.
- Hausmann, R., van de Sande, B., & VanLehn, K. (2008). Shall we explain? Augmenting learning from intelligent tutoring systems and peer collaboration. In B. P. Woolf, E. Aimeur, R. Nkambou & S. Lajoie (Eds.), *Intelligent tutoring systems* (pp. 636–645). New York: Springer.
- Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *The Physics Teacher*, 30(March), 141–158.
- Hickey, D. T., Kindfield, A. C. H., Horwitz, P., & Christie, M. A. T. (2003). Integrating curriculum, instruction, assessment, and evaluation in a technology-supported genetics learning environment. *American Educational Research Journal*, 40(2), 495–538.
- Horwitz, P., Gobert, J., Buckley, B. C., & Wilensky, U. (2007). *Modeling across the curriculum annual report to NSF*. Concord, MA: The Concord Consortium.
- Hmelo-Silver, C. E., Jordan, R., Liu, L. Gray, S., Demeter, M., Rugaber, S., Vattan, S., & Goel, A. (2008). Focusing on function: Thinking below the surface of complex science systems. *Science Scope*. Retrieved from <http://dilab.gatech.edu/publications/Science-Scope-Paper.pdf>
- Ioannidou, A., Repenning, A., Webb, D., Keyser, D., Luhn, L., & Daetwyler, C. (2010). Mr. Vetro: A collective simulation for teaching health science. *International Journal of Computer-Supported Collaborative Learning*, 5(2), 141–166.
- Klopfer, E., & Squire, K. (2008). Environmental detectives—the development of an augmented reality platform for environmental simulations. *Educational Technology Research & Development*, 56(2), 203–228.
- Kopriva, R. (2000). *Ensuring accuracy in testing for English language learners*. Washington, DC: Council of Chief State School Officers.
- Krajcik, J., Marx, R., Blumenfeld, P., Soloway, E., & Fishman, B. (2000). *Inquiry-based science supported by technology: Achievement and motivation among urban middle school students*. Paper presented at the American Educational Research Association, New Orleans, LA.
- Lai, C. Y., & Wu, C. C. (2006). Using handhelds in a jigsaw cooperative learning environment. *Journal of Computer Assisted Learning*, 22(4), 284–297.
- Lehrer, R., Schauble, L., Strom, D., & Pligge, M. (2001). Similarity of form and substance: Modeling material kind. In D. K. S. Carver (Ed.), *Cognition and instruction: 25 years of progress* (pp. 39–74). Mahwah, NJ: Lawrence Erlbaum Associates.
- Lim, K. Y. T., & Wang, J. Y. Z. (2005). Collaborative handheld gaming in education. *Educational Media International*, 42(4), 351–359.

- Liu, T.-Y., Tan, T.-H., & Chu, Y.-L. (2009). Outdoor natural science learning with an rfid-supported immersive ubiquitous learning environment. *Journal of Educational Technology & Society*, 12(4), 161–175.
- Maehr, M. L. & Midgley, C. (1996) *Transforming school cultures*. Boulder, CO: Westview Press.
- Martin, J., & VanLehn, K. (1995). Student assessment using Bayesian nets. *Journal of Human-Computer Studies*, 42(6), 575–591.
- McDowell, L., Sambell, K., Bazin, V., Penlington, R., Wakelin, D., Wickes, H., & Smailes, J. (2006). *Assessment for learning: Current practice exemplars from the Centre for Excellence in Teaching and Learning in Assessment for Learning*. Available at http://www.northumbria.ac.uk/sd/central/ar/academy/cetl_afl/pubandpres/intpub/occasional/
- Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, 32, 13–23.
- Minstrell, J., & Kraus, P. A. (2007). *Applied research on implementing diagnostic instructional tools*. Seattle, WA: FACET Innovations.
- Mislevy, R. J., & Gitomer, D. H. (1996). The role of probability-based inference in an intelligent tutoring system. *User-Modeling and User-Adapted Interaction*, 5, 253–282.
- Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessment (with discussion). *Measurement: Interdisciplinary Research and Perspective*, 1(1), 3–62.
- Morrison, A., Oulasvirta, A., Peltonen, P., Lemmela, S., Jacucci, G., Reitmayr, G., . . . Juustila, A. (2009). *Like bees around the hive: A comparative study of a mobile augmented reality map*. Paper presented at the 27th international conference on human factors in computing systems.
- Nersessian, N. J. (2008). *Creating scientific concepts*. Cambridge, MA: MIT Press.
- Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education*, 31(2), 199–218.
- Norman, D.A. (1993). *Things that makes us smart*. Reading, MA; Addison-Wesley.
- NRC. (2009). *Learning Science in Informal Environments: People, Places, and Pursuits*. Washington, DC: The National Academies Press.
- NRC. (2010). *Surrounded by Science: Learning Science in Informal Environments*. Washington, DC: The National Academies Press.
- Oberholzer-Gee, F., Waldfoegel, J. & White, M. (2010). Friend or foe? Cooperation and learning in high-stakes games. *The Review of Economics and Statistics*, 92(1), 179–187.
- Palincar, A. D., Brown, A. L. & Campone, J. (1991). Dynamic assessment. In H. L. Swanson (Ed.), *Handbook on the assessment of learning disabilities: theory, research, and practice* (pp. 75–95). Austin, TX: PRO-Ed.
- Patten, B., Sanchez, I. A., & Tangey, B. (2006). Designing collaborative, constructionist and contextual applications for handheld devices. *Computers & Education*, 46, 294–308.
- Pellegrino, J., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.

- Quellmalz, E. S. (2009). Assessing new technological literacies. In F. Scheuermann & F. Pedro (Eds.), *Assessing the effects of ICT in education: Indicators, criteria, and benchmarks for international comparisons*. Luxembourg: European Union/OECD.
- Quellmalz, E. S., DeBarger, A., Haertel, G., & Kreikemeier, P. (2005). *Validities of science inquiry assessments: Final report*. Menlo Park, CA: SRI International.
- Quellmalz, E. S., & Haertel, G. (2004). *Technology supports for state science assessment systems*. Paper commissioned by the National Research Council Committee on Test Design for K–12 Science Achievement.
- Quellmalz, E. S., & Haertel, G. D. (2008). Assessing new literacies in science and mathematics. In J. D. J. Leu, J. Coiro, M. Knowbel, & C. Lankshear (Eds.), *Handbook of research on new literacies*. Mahwah, NJ: Erlbaum.
- Quellmalz, E. S., & Moody, M. (2004). *Models for multi-level state science assessment systems*. Paper commissioned by the National Research Council Committee on Test Design for K–12 Science Achievement.
- Quellmalz, E. S., & Silberglitt, M. S. (2010). *Integrating Simulation-Based Science Assessments into Balanced State Science Assessment Systems*. Paper presented at the annual meeting of the American Educational Research Association, Denver, CO.
- Quellmalz, E. S., Timms, M. J., & Buckley, B. C. (2009). *Using science simulations to support powerful formative assessments of complex science learning*. Paper presented at the American Educational Research Association, San Diego, CA.
- Quellmalz, E. S., Timms, M., & Buckley, B. C. (in press). Exploring the role of technology-based simulations in science assessment: The Calipers project. *International Journal of Learning Technologies*.
- Rieber, L. P., Tzeng, S., & Tribble, K. (2004). Discovery learning, representation, and explanation within a computer-based simulation. *Computers and Education*, 27(1), 45–58.
- Sambell, K. (2010). Enquiry-based learning and formative assessment environments: Student perspectives. *Practitioner Research in Higher Education*, 4(1), 52–61.
- Sausner, R. (2004). Ready or not. District administration. Retrieved May 29th, 2007 from <http://districtadministration.ccsct.com//page.cfm?p=832>
- Schwartz, D. L., & Heiser, J. (2006). Spatial representations and imagery in learning. In R. K. Sawyer (Ed.), *The Cambridge handbook of the learning sciences*. Cambridge: Cambridge University Press.
- Schmitt, B., Bach, C., Dubois, E., & Duranthon, F. (2010). *Designing and evaluating advanced interactive experiences to increase visitor's stimulation in a museum*. Paper presented at the 1st Augmented Human International Conference.
- Schultz, P., & Shugart, E. (2007). *Communicating with museum audiences about climate science: Contributions of Gene Rasmusson*. Presented at the Gene Rasmusson Symposium.
- Slotta, J. D. & Chi, M. T. H. (2006). The impact of ontology training on conceptual change: Helping students understand the challenging topics in science. *Cognition and Instruction* 24(2), 261–289.

- Squire, K. D., & Jan, M. (2007). Mad city mystery: Developing scientific argumentation skills with a place-based augmented reality game on handheld computers. *Journal of Science Education and Technology*, 16(1), 5–29.
- Sternberg, R. J. & Grigorenko, E. (Eds.). (2002). *Dynamic testing*. Cambridge, UK: Cambridge University Press.
- Stewart, J., Cartier, J. L., & Passmore, C. M. (2005). Developing understanding through model-based inquiry. In M. S. Donovan & J. D. Bransford (Eds.), *How students learn* (pp. 515–565). Washington, DC: The National Academies Press.
- Stewart, I., & Golubitsky, M. (1992). *Fearful symmetry: Is God a geometer?* Cambridge, MA: Blackwell Cambridge.
- Stieff, M., & Wilensky, U. (2003). Connected chemistry—incorporating interactive simulations into the chemistry classroom. *Journal of Science Education and Technology*, 12(3), 285–302.
- Stiggins, R. (2006). Assessment for Learning: A key to motivation and achievement. *Edge*, 2(2), 3–19.
- Timms, M. J. (2007). *Using item response theory (IRT) to select hints in an ITS*. Paper presented at the Artificial Intelligence in Education, Marina del Ray, CA.
- Twing, J. S. & Dolan, R. P. (2008). *UD-CBT guidelines*. Retrieved March 5th, 2008 from <http://www.pearsonedmeasurement.com/cast/index.html>
- Vonderwell, S., Sparrow, K., & Zachariah, S. (2005). Using handheld computers and probeware in inquiry-based science education. *Journal of the Research Center for Educational Technology*, 1(2), 1–11.
- Vygotsky, L. S. (1987). *The collected works of L.S. Vygotsky. Vol I*. New York: Plenum.
- Wang, S. (2005). *Online or paper: Does delivery affect results?* San Antonio, TX: Harcourt Assessment, Inc.
- Weller, A. M., Bickar, J. C., & McGuinness, P. (2008). Making field trips podtastic! Use of handheld wireless technology alleviates isolation and encourages collaboration. *Learning & Leading with Technology*, 35(6), 18–21.
- White, B. Y., & Frederiksen, J. R. (1998). Inquiry, modeling, and metacognition: Making science accessible to all students. *Cognition and Instruction*, 16(1), 3–118.
- Wilson, M. R., & Bertenthal, M. W. (Eds.). (2006). *Systems for state science assessment*. Washington, DC: The National Academies Press.

Author Queries:

- 1 – On p. 4, you cite Feuerstein, Rand & Hoffman, 1979, but this is not included in your reference list. Please add this item to your references.
- 2 – You include Sternberg & Grigorenko, 2002 in your references, but this is not cited in the text. Please add a citation or remove this item from your reference list.
- 3 – On p. 3 and again on p. 12 you cite Buckley (in preparation), but you do not specify which one of the two Buckley items you are citing. Please add an “a” or “b” to each citation.
- 4 – Please provide page numbers for the Quellmalz, E. S., & Haertel, G. D. (2008) chapter.

