21st Century Science Assessments

Edys Quellmalz, Jodi Davenport, & Mike Timms
WestEd

Introduction

Technology-based science tests are rapidly appearing in state, national, and international testing programs. For example, the Science Framework for the 2009 National Assessment of Educational Progress (NAEP) calls for the design of computer-based science investigation tasks. The 2006, 2009, and 2012 cycles of the Programme for International Student Assessment (PISA) include computer-based forms (National Assessment Governing Board [NAGB], 2008; Koomen, 2006). The 2014 NAEP for Technology and Engineering Literacy will also administer interactive scenario-based tasks. At the state level, Minnesota has an online science test with simulated laboratory experiments and investigations of phenomena such as weather or the solar system (Minnesota Department of Education, 2010). Utah is piloting science simulations in their assessments (King, 2011). The state testing consortia are designing technology-enhanced items to test English Language Arts and Math common core standards, so it is likely that tests of the forthcoming Next Generation Science Standards will include innovative task and item formats.

The powerful capabilities of technology may hold the key to transforming both the science knowledge and practices that are tested and how they are tested (Quellmalz, et al., 2011) Quellmalz & Pellegrino, 2009; Quellmalz & Haertel, 2004). Technologies allow representation of domains, systems, and data in new ways and support novel forms of interactions. Scientists use physical, mathematical, and conceptual models as tools for generating and testing hypotheses and to communicate about natural and designed systems (Nersessian, 2008; Clement, 1989).

Unfortunately, the press to implement more complex computer-based science tests has outpaced research addressing crucial questions about the validity, comparability, and complementarities of new science assessment task designs. Also, while most test developers are highly experienced in creating static items, developers are less likely to be aware of findings from learning research that have studied how students perform best in interactive multimedia formats. As assessment tasks and items become more interactive, developers need to take into account these bodies of cognitive and multimedia research that offer much that is relevant to the development of complex technology-based assessments. Moreover, deeper study of the affordances of technology-enhanced assessment designs can shed light on the benefits of different formats and modalities for measuring understanding of complex science system dynamics and inquiry practices.

Goals

This presentation summarizes findings from a research project funded by the National Science Foundation, Foundations of 21st Century Science Assessments, which studied the design and construct validity of science assessments that use static, dynamic, and interactive task design structures to test challenging science standards. The specific goals of the project were to:

- Derive research-based principles from the literature for designing next generation science assessment tasks and items that draw on the capabilities of technology to elicit understanding of science systems and inquiry abilities.
- Examine existing large-scale science tests to determine the extent to which they: (1) address science standards for understanding science systems and using inquiry practices that are set forth in national science education frameworks and standards, and (2) employ research-based design principles.
- Investigate the validity of technology-based assessment tasks and items that vary in levels of interactivity (static, active, and interactive) for measuring the constructs of science inquiry practices.

Theoretical Framework

The project surveyed research on learning and assessment to distill design principles for next generation science assessments. The major bodies of research included studies of cognition, model-based learning, learning from simulations, multimedia learning, and measurement methodology. The assessments subsequently designed to compare modalities were designed according to recommendations for quality science assessment being made by cognitive scientists, learning theorists, science educators

Model-based Learning

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 models, or schema, of system dynamics and components (Nersessian, 2008). Cognitive research shows that learners who internalize schema of complex system organization—structure, functions, and emergent behaviors—can transfer this heuristic understanding across systems (cf., Goldstone, 2006; Goldstone & Wilensky, 2008). This body of research, coupled with the call of the new *Framework for K-12 Science Education* for integrating science content into systems thinking and more attention to the practices of science, suggest that the next generation of science assessments should include task designs that model the organizational schema characteristic of all complex systems--components, interactions, and emergent behaviors and incorporate investigations to assess inquiry practices.

Research on Learning in Science

If students have a deep understanding of a science system they should understand core concepts and be able to use their knowledge to make inferences and conduct scientific investigations. Thus, the challenge of science assessment is to develop tasks that do not simply tap into disconnected bits of declarative and procedural knowledge, but that call for the schematic and strategic knowledge needed to reason about complex systems and engage in scientific inquiry practices. Guidelines for designing assessment tasks that test inquiry practices and model-based reasoning can be derived from numerous publications and reports that synthesize decades of research on human learning and strategies for promoting its development, such as presentation of meaningful tasks, opportunities for active inquiry, individualized scaffolding, and opportunities for self assessment and scientific discourse.

Affordances of Simulation-based Environments

For the next generation science assessments, simulations can become a resource for presenting authentic, complex tasks. Simulations can represent models of science systems—components, interactions and emergent behaviors—and dynamically portray spatial, temporal, and causal phenomena. Simulations permit manipulations of alternative system conditions to investigate predictions. Moreover, as simulations can superimpose multiple physical and symbolic representations, they can reduce potentially confounding language demands.

Multimedia Learning Research

Though visualizations and simulations have many affordances for learning, the additional information they present may also distract or overwhelm students. Multimedia learning researchers have examined the effects of pictorial and verbal stimuli in static, animated, and dynamic modalities, as well as the effects of active versus passive learning enabled by degrees of learner control (Clark & Mayer, 2011; Mayer, 2005; Lowe & Schnotz, 2007). The majority of multimedia design principles address how to focus students' attention and minimize extraneous cognitive processing.

Animations are considered particularly useful for providing visualizations of dynamic phenomena that are not easily observable in real space and time scales, cf., plate tectonics, circulatory system, animal movement, (Betrancourt, 2005; Kuhl, Scheiter, Gerjets & Edelmann, 2011). User control can allow students to replay dynamic presentations and can increase the likelihood that students will learn from the display (cf., Lowe & Schnotz, 2008; Schwartz & Heiser, 2005).

Evidence-Centered Assessment Design

The NRC report, *Knowing What Students Know,* integrated the learning research summarized in *How People Learn* with advances in measurement science to describe systematic test design frameworks. Evidence-centered assessment design involves relating the learning to be assessed, as specified in a *student model*, to a *task model* that specifies features of the task and questions that would elicit observations of learning, to an *evidence model* that specifies what student responses and scores would serve as evidence of proficiency (Messick, 1994; Mislevy et al., 2003; Pellegrino et al., 2001).

Cognitively principled assessment design for science begins with a student model derived from a theoretical framework of the kinds of knowledge structures and strategies students should demonstrate as evidence of their level of expertise. The model-based learning and national science education frameworks and standards identify the broad conceptual knowledge structures and inquiry practices deemed by the profession to be goals of science education (College Board, 2009; NAGB, 2008; NRC, 2011; AAAS, 1993). For this study, the model-based learning framework guided the representation of science system phenomena into the three levels of components, interactions, and emergent system behavior. The science practices set forth for the 2009 Science NAEP guided the science inquiry targets specified in the student model and informed the task model.

The science practices and their cognitive demands referenced the 2009 NAEP Science Framework included:

Identifying principles - declarative knowledge, cognitive demand "knowing that"

- Knowing components and their rules
- Recognize principles, definitions

Using principles – declarative and schematic knowledge, integrated knowledge structures, cognitive demand "knowing why"

- Use patterns in observations
- Making predictions (What would happen if…)
- Creating explanations

Conducting Inquiry – procedural and strategic knowledge, cognitive demand "knowing how and when"

- Designing experiments
- Testing predictions
- Evaluating explanations

Specifications of the evidence models were based on identifying the types of student responses within the simulation-based tasks that would serve as evidence of proficiency on science knowledge related to the three system model levels and to the inquiry practices specified in the NAEP 2009 Framework. Rules were generated for scoring responses and summarizing them in order to report the proficiency levels.

The task models integrated the system model levels and the inquiry practices for the levels and shaped the types of authentic, problem based tasks, their sequences, and specific stimulus and response features of the tasks in the static, active, and the interactive modalities.

Methods

Research questions

The study addressed the following two questions:

1. How well do extant items represent research-based design principles and inquiry practices?
2. Do student responses on assessments with different modalities (static, active, and interactive) provide different information about students' proficiencies on the dimensions of science practices: knowledge of science principles, use of science principles, and ability to conduct scientific inquiry?

Study design

The study proceeded in three phases. In Phase 1, principles for designing next generation science assessments were culled from the literature review. In Phase 2, the project collected a sample of items from existing large-scale science tests to analyze their coverage of science system model levels and inquiry practices and their adherence to research-based design principles. In Phase 3, the project created parallel sets of static, active, and interactive tasks intended to test identical science inquiry practices constructs. Student performance on tasks and items in the three modalities were compared to determine if the different levels of interactivity permitted by each modality elicited distinctive evidence of performance on the inquiry practices.

Phase 1: Design principles.

For question 1, the project distilled a set of principles for constructing next generation science assessment learning targets and tasks from the learning studies cited earlier. For the purposes of this article, we describe the principles that apply to the design of the summative

assessments that were designed for this study to compare modalities. These principles are organized below as components of the evidence-centered assessment design framework.

Student Model. Employ a science system framework that integrates core science knowledge (science systems) and inquiry practices assessment targets.

Evidence Model. *Before and during* development of assessment tasks and items specify the evidence of knowledge and inquiry practices targets to be collected, as well as the scoring and reporting rules.

Task Model. Design assessment tasks and items that will elicit evidence of the knowledge and inquiry practices. The task designs should:

1. Present *meaningful, authentic, recurring problems.*
2. Require *active construction* of knowledge and use of inquiry/problem solving practices and metacognitive skills.
3. Use multiple tasks to assess mental models by probing multiple points in a system (*multiple probes).*
4. *Sequence* task components to gradually increase the complexity of the phenomena and inquiry practices.
5. Limit the task environment to the scope relevant to the problem at hand (*Fidelity and Coherence principles)*. Ensure the representations of the science phenomena are *necessary* and *sufficient* to respond to questions and accomplish tasks.
6. Make the most important information salient, omit or make less important information less salient (*Attention guiding principle*).
7. Put labels next to relevant pieces of information (*Contiguity*).
8. For dynamic modalities *(animations, simulations),* provide for *user control*. Students should have the ability to control the viewing and pace of the information and have the flexibility to return to transient information.
9. The level of interactivity should reflect the required task demands.
10. Use visual and textual elements to guide attention, parse complex animations or simulations into meaningful chunks, and give cues, (e.g., "There will be 3 steps.") (*Signaling and Segmenting*).

Phase 2: How do extant assessment items represent research-based design principles and inquiry practices?

Methods.

To determine whether existing assessments adhere to design principles and cover the range of science practices specified in the 2009 NAEP Science Framework, the study investigated the types of science system knowledge and inquiry skills addressed in existing middle school science assessments administered at state, national, and international levels.

Sample. The study searched for items related to two fundamental life and physical science topics, ecosystems and chemistry. The search resulted in item pool of released and sample items from more than 30 state, national, and international tests. Using alignment to practices in the 2009 NAEP Science Framework as the criterion for including items, 98 static items from 21 assessments were found that related to either ecosystems or chemistry content targets at the middle school level. Because of the emerging nature of the field, researchers located only six active items (from the Minnesota state science assessment and PISA). No interactive items related to ecosystems or chemistry targets at the middle school level were found in the publicly available items.

Measures. An instrument was developed to classify the items according to the types of science system knowledge assessed, the three NAEP 2009 science practices, and types of knowledge/cognitive demands. Two reviewers coded 98 static and the six active items. Roles, Interactions, and Populations. Similarly, chemistry items were judged in terms of their alignment to three targets—Components (Particles), Interactions (Microscopic Level), and Emergent Properties (Macroscopic Level). Reviewers coded items to one or more of the following NAEP science practices: *Identifying Principles*, *Using Principles*, Conducting Inquiry (i.e., designing investigations, conducting investigations, analyzing data, and drawing conclusions.

Additionally, reviewers analyzed the cognitive demand of items, coding them to one or more of the following categories: Declarative, Procedural, Schematic, and Strategic. Finally, reviewers analyzed items in terms of the design principles. This involved analyzing the representations (e.g., text, pictures, graphs, and animations) in terms of their role in the item (required, clarifying, or extraneous), clarity, grade-level appropriateness, complexity, and relevance, as well as whether visual cues supported mapping between representations.

Design and procedure. Two researchers used an online form and a companion coding document to code the existing items. Overall, the average pairwise percent agreement of 82.7% and Cohen's Kappa was calculated at 0.67, indicating substantial agreement. All discrepancies in coding were discussed and reconciled among the reviewers.

Results. The results reported here are based on analysis of the reconciled data.

Content. Table 1 summarizes the coding of the 98 static items (ecosystem 48) and (chemistry items 50). For the ecosystem items, most were judged to assess student knowledge of interactions (30) or components/roles (25). Reviewers only coded seven items to the emergence (populations) level. For the 50 chemistry items, 40 items were coded at the emergence level related to states of matter, while fewer items assessed knowledge of components (7) and interactions (9). Items that involved more than one level of the science system were dual coded.

Table 1. Distribution of items by model level.

	Components	Interactions	Emergence
number of Ecosystems items	25	30	7
number of Chemistry Items	7	9	40

Science Practices. The large majority of the 98 static items in the sample were judged to test the first two science practices associated with content knowledge—*Identifying Principles* (56) and *Using Principles* (63). Twelve items assessed the practices associated with *Conducting Inquiry* practices. Six of the *Conducting Inquiry* items addressed communicating findings, identifying patterns, or designing investigations. Only one item asked students to conduct investigations. The reviewers did not code any items to the NAEP science practice of drawing or evaluating conclusions.

Cognitive Demand. Only five of the 98 static items analyzed were judged to address strategic thinking, while 94 items called for declarative knowledge. Reviewers also coded 52 items as requiring schematic knowledge ("knowing why") and 18 items testing procedural knowledge ("knowing how"). Items could be coded to more than one cognitive demand.

Multiple representations and design principles. In general, the majority of items adhered to the design principles. All items contained text. Ninety-six percent of these items contained only relevant or related text and 84% of the items adhered to text-related design principles such as clarity and task appropriateness.

Items that contained other representations such as diagrams and graphs were slightly less likely to adhere to design principles. Of the 46 items that contained *diagrams*, 78% adhered to design principles. Only 4 items contained *graphs*, and only one of these items met the design principles.

Only six dynamic items (animations) were found that aligned with ecosystems content (4) or chemistry (2), two of the six animations were required for the task, while the other animations merely provided context for the storyline. All animations were coded to adequately represent temporal information, to be scientifically appropriate, and to present simultaneous auditory information.

Discussion

The analysis of items from existing large scale science assessments for two prominent science systems--ecosystems and chemistry—revealed dramatically uneven attention to knowledge about the three levels of science systems. The majority of the items required declarative knowledge rather than more complex cognitive processing. Moreover, the analyses added to findings from prior research that items on the current assessments do not tap some important science practices (Quellmalz, et al., 2005). Further, existing items that used pictorial information, such as diagrams or graphs, were less likely to adhere to the design principles derived from the literature, suggesting that next generation task and item designers will need to carefully formulate, combine, and study the multiple representations and their interrelationships. The analyses add support to concerns that large-scale assessments emphasize declarative and procedural knowledge, with fewer items requiring strategic thinking and reasoning skills. These data suggest a need for new designs for next generation tasks that test knowledge of science system models and inquiry practices.

Phase 3. Comparison of the three assessment modalities.

The main question in our study was whether or not student responses on assessments in different modalities (static, active, and interactive) provide different information about students' proficiencies on the science practice constructs. To compare how well assessments in different modalities were able to measure the science practice constructs, we developed three parallel assessments in the context of the life science topic of ecosystems. Items in the three modalities were designed to test the same science practice constructs and to be comparable on all key stimulus and response features—except for dynamic representations of science phenomena and degrees of interactivity**,** which varied across the static, active, and interactive modalities. We then analyzed our data using multiple psychometric techniques to determine whether the dynamic, active, and interactive assessments were better able to independently estimate student performance across the three science practices (*Identifying Principles, Using Principles,* and *Conducting Inquiry*).

Method.

Participants. A total of 1,836 students (910 female, 899 male, and 27 of unrecorded gender) from the classrooms of 22 middle school science teachers in 12 states participated in the study as part of normal classroom activities. Teachers received a stipend for the time needed to

complete study activities (e.g., providing demographic information and enrolling students in the online learning management system). Due to absences, only 1566 students (778 female, 776 male, 12 unknown) completed all three versions of the assessment. Thus, the total sample size included in the analyses was 1,566.

Materials. For each of the three modalities, six items were designed to test the construct of *Identifying Principles*, six items were designed to test *Using Principles*, and 12 items were designed to test *Conducting Inquiry* for a total of seventy-two items in the context of ecosystems. As the interactive modality for ecosystems had been developed and validated in prior research, the tasks aligned with the science practice targets served as the starting point for the design of the other two modalities (Quellmalz, et al., 2010; Quellmalz, et al., 2011). The existing ecosystem simulation environments and templates for ecosystems model levels and inquiry tasks were used to generate the parallel item sets.

The three modalities were set within three different ecosystems, (tundra, grasslands, and mountain lake). In the static modality, students viewed still images on the screen within a tundra ecosystem. In the active modality, students viewed animations of a grasslands ecosystem, but did not manipulate features or conduct active investigations. In the interactive modality, students identified and used ecosystem principles within a mountain lake ecosystem and conducted inquiry in tasks such as designing and running their own experiments. Each modality assessed the same science inquiry practice constructs. Figure 1 illustrates the science constructs of *Identifying Principles* and *Using Principles* in a food web task in the different modalities. In the static modality, students read about the interactions in an ecosystem and were asked to select the static image of the correct food web. In the active modality, students observed an animation of an ecosystem to infer organism roles and then drew a food web diagram. In the interactive modality, students observed an ecosystem and could take advantage of additional interactivity that cued the connection between the names and pictures of the organisms.

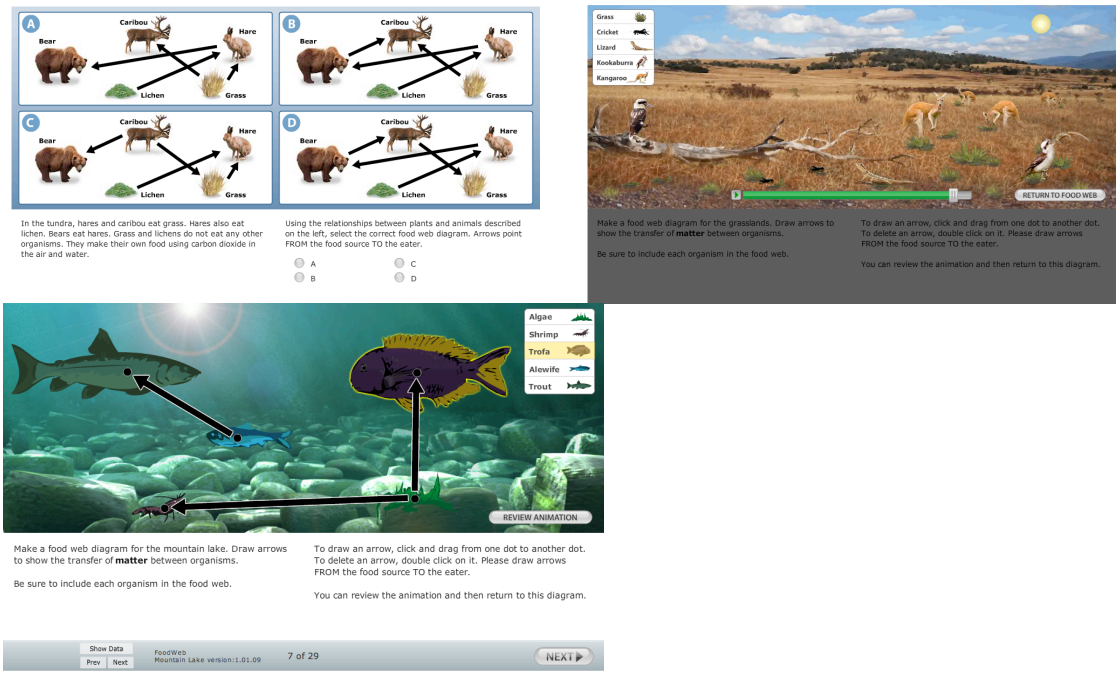

Figure 1. Ecosystem Food Web task in static (top left), active (top right), and interactive modalities (bottom).

Figure 2 illustrates items designed to test the construct *Conducting Inquiry* in the different modalities. In the static modality, students viewed the outcomes of an experiment and were asked to select an appropriate evaluation of the design. In the active modality, the student watched an animation of the graphs being generated, but could not set the slider values. Finally, in the interactive modality, students set the slider values, then ran and saved their own trials.

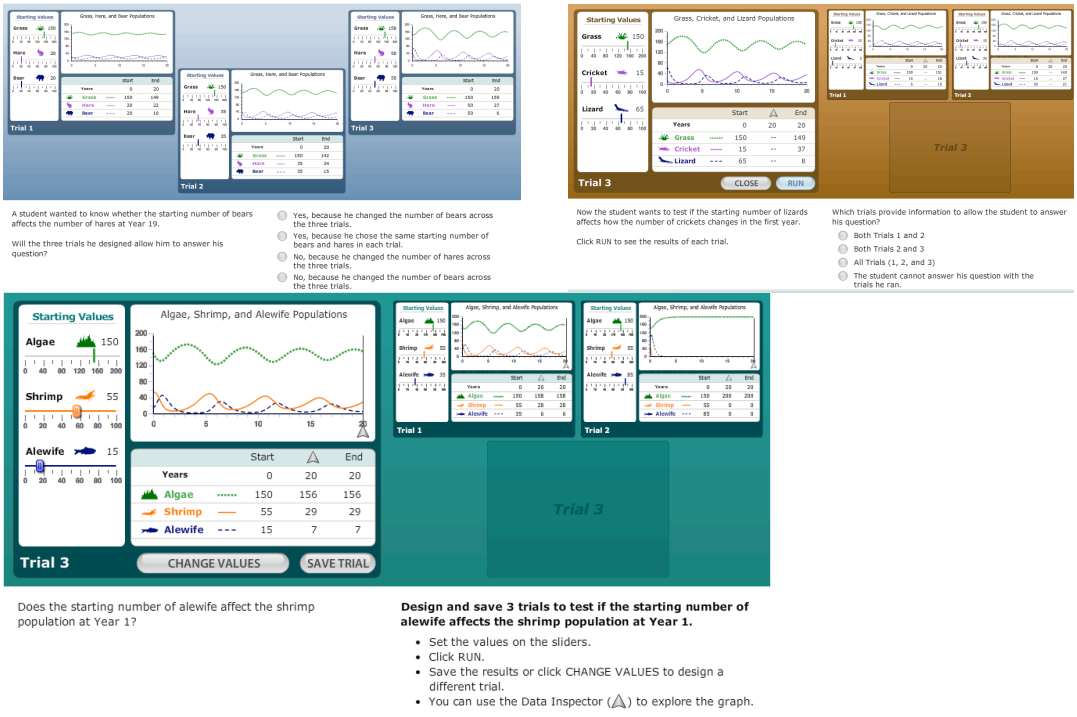

Figure 2. Experimental design task in static (top left), active (top right), and interactive modalities (bottom).

Design and procedures. All items were administered online and data were collected using the SimScientists Learning Management System (Quellmalz, Timms, Silberglitt, & Buckley, 2011). The learning management system allows teachers to check on individual students and researchers to download de-identified student data. Construct validity of the items was examined initially by expert reviews and cognitive labs.

Expert reviews. At two points in the process of developing the parallel item sets, three experts from the American Association for the Advancement of Science (AAAS) independently reviewed the items and judged if each item was aligned with one of the targeted science practices of Identifying Principles, Using Principles or Conducting Inquiry. These experts also verified that the items were scientifically accurate, grade-level appropriate, usable, and comparable across the static, active and interactive versions. Initially, AAAS staff reviewed the storyboards of draft items and provided detailed comments and feedback. An additional iteration of review and revision was carried out with the programmed items to ensure the final items remained aligned with targeted science inquiry practices.

Cognitive Labs. Ten students participated in think aloud studies to determine if the items elicited the targeted science inquiry practices. Each student completed all three forms of the assessments, one in each modality (static, active, interactive). As students completed the assessments, they "thought aloud" by saying everything they were thinking while screen capture software recorded students' verbalizations and actions on the screen and researchers coded whether the items elicited the targeted construct. The think-aloud studies had two goals: 1) to ensure the usability of the assessments as deployed, and 2) to provide evidence of construct validity by determining that the questions were eliciting student thinking and reasoning about the intended science practice constructs. To ensure the items would be usable in the field test,

researchers took detailed notes of usability issues that arose (e.g., navigation, difficulty running experimental trials) and modified the items to address these issues. To examine the items' construct validity, the observing researcher coded whether the item prompted student thinking related to the targeted science practice constructs. Table 2 summarizes the percentage of items in the assessments judged to elicit their intended construct targets. These data provided one form of evidence that the items were aligned with their intended content and inquiry targets.

Table 2. Percentage of items judged to elicit their intended construct targets.

	Identifying principles	Using principles	Conducting inquiry
Static	100%	97%	98%
Active	98%	98%	100%
Interactive	98%	98%	98%

The study used a within-subjects design, as all participating students took all three versions of the ecosystems assessments, i.e, one period of static items, one period of active items and one period of interactive items. The assessments were given in three consecutive sessions and the order of the sessions was fully counterbalanced at the class level across the six possible sequences of assessments: static (S), active (A), interactive (I), SIA, ISA, IAS, AIS, ASI). Table 3 shows the number of items in each test session.

Table 3. Number of items in each test session by modality and science practice construct.

		Test Session A	Test Session B	Test Session C	
		Static Modality	Active Modality	Interactive Modality	Total Items
Science Practice Constructs	Identifying Science Principles	6	6	6	18
	Using Science Principles	6	6	6	18
	Conducting Science Inquiry	12	12	12	36
Total Items		24	24	24	72

Participating teachers received a detailed "Teacher Guide" that outlined step-by-step the processes and procedures for study activities and were able to view online movies that demonstrated how to carryout the online processes and procedures necessary for participation. During the study, students either used laptops in their science classrooms or went to the school computer lab.

Analyses

In order to answer the second research question, we compared how well each assessment modality measured the three science practice constructs using three different types of analyses: (1) a generalizability study (G-study), (2) a Multitrait-Multimethod Confirmatory Factor

Analysis and (3) a multidimensional Item Response Theory (IRT) model. Each takes a slightly different approach to modeling the data, as explained below.

Generalizability study.

A G-study was chosen as the first analytic method because it allows a quantitative investigation of how much error in a data set is attributable to different facets (sources of variation) and the interactions among facets. In this study the facets were the persons (students) and the items.

Two sets of multivariate G-study analyses were conducted using the mGENOVA computer program (Brennan, 2001). Given that there were three tasks (Grasslands, Tundra and Mountain Lake) corresponding to three modalities of item format (static, active and interactive) and that each task consisted of items addressing three different science practice constructs (identifying, using, conducting), the first set of analyses was a multivariate G-study treating the nine modality x construct combinations as nine separate constructs. In the mGENOVA terminology, this is a $p^{\bullet}$ x i^{o} design. The p facet represents persons (students). The solid circle means that the same students responded across all nine constructs. The i represents items. The open circle means that the items were nested within each of the nine constructs. This $p^{\bullet}$ x i^{o} notation represents the univariate design for each of the nine separate constructs. An additional "v" facet was used for the nine constructs.

The assessments in the three modalities were designed to tap into the same science constructs. As the first set of analyses ignored the linkage of items across the three modalities in a given construct, a second set of analyses (one for each of three constructs) was conducted to address the links of items across assessments. In each analysis, the fixed facet had just three levels, corresponding to the three modalities. Both the person facet and the item facet were linked across levels of the fixed facet. Therefore, this was a $p^{\bullet}$ x $i^{\bullet}$ design.

Multitrait-Multimethod Confirmatory Factor Analysis (CFA).

The second type of analysis selected for answering the research questions was a multitrait-multimethod CFA analysis (Campbell & Fiske, 1959; Loehlin, 1998), which attempts to separate out the true variance on measured traits from the variance that is due to the method of measurement. It is well-suited to this study because the same traits/constructs (the three science practices) were measured with three different methods (the static, active and interactive modalities). The resulting correlations among the different measurements were then arranged in a multitrait-multimethod matrix in order to assess the *convergent validity*, the tendency for different measurement methods to converge on the same trait/construct, and the *discriminant validity*, the ability to discriminate among different traits/constructs.

This analysis was used to test the following two hypotheses that stem from the second research question:

1. The factor loadings for method (assessment modality) are higher than for trait (science practice constructs).
2. The factor loadings from assessment modality to the three science practice constructs
    a. are less for the interactive modality than for the static and active modality OR
    b. the correlations between constructs are generally smaller for the interactive modality than for the static and active modalities.

The multitrait-multimethod analysis used the same data set of 1,566 complete responses as was used for the G-study analyses. However, instead of using the responses to 72 individual items, nine composite scores were computed and used in the CFA. This was to simplify the analysis and the interpretation and to reduce the possibility of the estimation not being able to converge during analysis. The nine composite scores were simply the sum of the 6 items for *Identifying Principles* and *Using Principles* items or the sum of 12 items for *Conducting Inquiry* items, computed for each of the three modalities (static, active and interactive). Two separate analyses were run. First, a model with three factors was fitted to the nine scores, with each score permitted to load only on one factor representing the Science Practice construct measured (Identifying, Using, or Conducting). Second, a model with three factors was fitted to the nine scores, with each score permitted to load only on one factor representing its Assessment Modality (Static, Active, or Interactive). The Mplus[1] computer program was used for this analysis.

Multidimensional Item Response Theory Model.

The third type of analysis used multidimensional IRT models to evaluate how well each assessment modality (static, active or interactive) was able to measure and separate student performances on the three sciences practice constructs. IRT models are probabilistic models in which item difficulty (a test item's underlying difficulty based on the proportion of a given sample that responded correctly) and person measure (a person's underlying competence, based on the proportion of items completed correctly) are simultaneously estimated. The result is a scale on which both persons and items are mapped onto the theoretical latent traits, which in this case are the science practices constructs. The fact that IRT scores' accuracy and precision can be quantified makes this a suitable analytic method in this study to determine how well each of the three modes of assessment measure the three science practice constructs.

The ACER Conquest[2] generalized item response modeling program was used to run a multidimensional logistic model analysis that modeled the three science practice constructs separately for each of the three modalities of assessment. This allowed the correlations among the three science practice constructs to be estimated and for the reliability of the measurement of each of the practice constructs to be quantified for the three modalities of assessment.

Results

G-study

Tables 3a-3c, which were produced from the first set of analyses in the G study, summarize the estimated correlations among the three science practice constructs for each of the three modalities of assessment. The estimated correlations are corrected for attenuation due to unreliability, i.e., estimated correlations among universe scores (true scores) for the nine constructs.

Table 3a. Estimated correlations among the three science practices for the static mode (Tundra)

	Identifying	Using	Conducting
Identifying	1	0.92	0.80
Using		1	0.91

[1] http://www.statmodel.com/

[2] ACER Conquest: Generalized Item Response Modeling Software published and distributed by the Australian Council for Educational Research

Table 3b. Estimated correlations among the three science practice constructs for the active mode (Grasslands)

	Identifying	Using	Conducting
Identifying	1	**0.80**	0.80
Using		1	1
Conducting			1

Table 3c. Estimated correlations among the three science practice constructs for the interactive mode (Mountain Lake)

	Identifying	Using	Conducting
Identifying	1	0.82	**0.72**
Using		1	**0.84**
Conducting			1

These correlation coefficients can be interpreted to indicate which modality of assessment is measuring the science practice constructs distinctly. While we expect there to be a positive correlation among the three individual science practice constructs because they are related elements of the overall set of science practices, if they are clearly observable skills, then the correlations will not be too high. If, for example, the correlations are .90 and above, then one could conclude that the measures are really only measuring a single construct and it would not be possible to make inferences about students' skills in each of the individual practice constructs.

For ease of reference, the lowest correlation for each pair of science practice constructs is shown in bold in the table. Two of the three lowest correlations among the three science practice constructs (.72 for *Identifying/Conducting* and .84 for *Using/Conducting*) are from the interactive assessment (Mountain Lake) modality. The active (Grasslands) assessment modality produced the lowest correlation of .80 for *Identifying/Using*, which was slightly lower than the .82 for the interactive modality. Conversely, two of the highest correlations (.92 *Identifying/Using* and .91 *Using/Conducting*) are for the static modality (Tundra), and the highest correlation (1.0 *Using/Conducting*) is for the active modality (Grasslands).

Overall, the results from the first analysis in the G Study suggest that the interactive modality measured the science practice constructs more distinctly and was particularly able to distinguish *Conducting Inquiry* as a clear construct, which the static and active modalities measured poorly.

Table 4 summarizes the second analysis in the G study. It presents the estimated variance components and G-coefficient by construct and indicates the percentage of variance attributable to each. For example, under the static assessment modality for the *Identifying Principles* construct, about 20% of variance is contributed by persons (students), about 16% is from items, and 64% (the majority of variance) is accounted for by the person x item interaction. In general, this pattern holds for each construct.

Table 4. Estimated variance components and G-coefficient by construct.

	Identifying			Using			Conducting		
Effect	Static	Active	Interactive	Static	Active	Interactive	Static	Active	Interactive

persons									
estimated variance	0.04	0.05	0.05	0.04	0.03	0.03	0.03	0.03	0.05
percentage of total	19.62	22.27	20.67	16.48	14.83	14.79	11.97	11.59	18.11
items									
estimated variance	0.04	0.04	0.04	0.01	0.02	0.03	0.05	0.04	0.06
percentage of total	15.93	14.64	15.15	5.96	7.17	12.30	19.92	16.33	22.49
person x item interaction									
estimated variance	0.15	0.15	0.16	0.17	0.17	0.17	0.17	0.17	0.15
percentage of total	64.45	63.09	64.18	77.56	78.00	72.91	68.11	72.08	59.40
G-Coefficient	0.65	**0.68**	0.66	**0.56**	0.53	0.55	0.68	0.66	**0.79**

Note: The highest correlations for each science practice are shown in bold.

The bottom row of Table 4 shows the G-coefficient estimate (a reliability-like coefficient) for each science practice construct measured under each of the three assessment modalities. For ease of interpretation, the highest correlations for each of the three science practice constructs are shown in bold. As explained earlier, the G-coefficients for the *Identifying* and *Using* constructs are based on a 6-item test within each modality, whereas the coefficients for *Conducting* are based on a 12-item test (of multiple component skills) within each modality. It should also be noted that the percentages in the body of the table describe the contributions to the variance of a single item, whereas the G coefficients pertain to composite scores consisting of either 6 or 12 items.

As in the first G-study analysis, the *Conducting Inquiry* construct is measured most reliably (.79) by the interactive modality. The reliability of the *Using Principles* construct was generally low with coefficients of between .53 and .56 and for the *Identifying Principles* construct coefficients ranged from .65 to .68. The differences were small considering that they were based on only 6 items and it is not possible to say that one assessment modality was clearly different from the others for measuring the *Identifying Principles* and *Using Principles* constructs.

Confirmatory Factor Analysis/Multitrait-multimethod

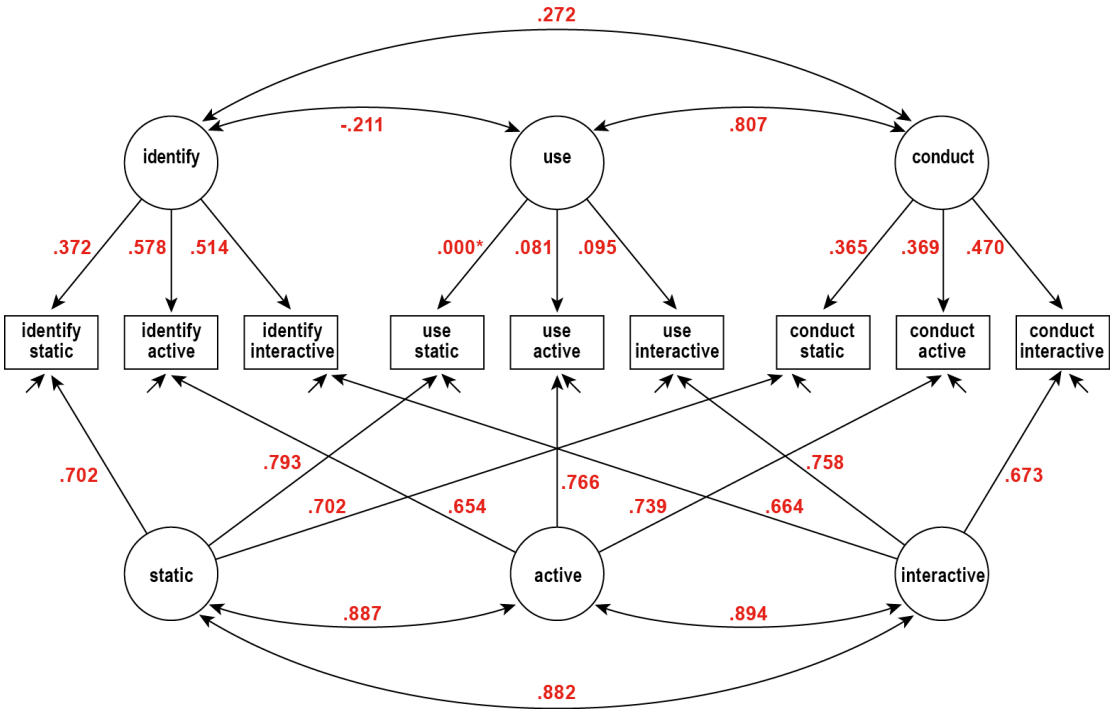

Figure 3. Path model for the CFA/Multitrait-multimethod matrix

Figure 3 shows the path model for the CFA/Multitrait-multimethod matrix analysis. In the diagram, the top three circles represent the factors included in the model for the three science practices (identify, use and conduct). The bottom three circles represent the model factors for the three assessment modalities (static, active and interactive). The nine rectangles represent the sets of items. The first three on the left represent the items that targeted the *Identifying Principles* science practice, and from left to right they represent the active, static and interactive item sets. The middle three rectangles represent the *Using Principles* items, again with static, active and interactive running from left to right. The final three rectangles on the right represent the *Conducting Inquiry* items, arranged yet again with static, active and interactive modalities running from left to right. The straight lines represent the loadings of items onto factors, with the standardized value represented by the figure against the line. The curved arrows represent the correlations between factors, with the value on each arc.

The first thing to notice about the factor loadings is that they are generally higher for the assessment modalities for the science practices, which indicates that the measurements (scores) are more determined by the assessment modalities than by the science practice constructs. By further looking at the factor loadings in Table 5, we also find that the factor loadings from the test modality to the three constructs are slightly less for the interactive modality (.698 on average) than for the static (.732) and active modalities (.720). This is supportive of our hypothesis that the task items in the interactive modality measure the science practice constructs more distinctly than the other two modalities. In particular, looking at the factor loadings in the

column for the *Conducting Inquiry* construct in Table 5, the loading for the interactive modality (.673) is considerably lower than for the static (.702) and active (.739) modalities. This is similar to the finding in the G-study that the interactive modality more distinctly measures the *Conducting Inquiry* construct.

Table 5. Assessment Mode Factor Loadings

	Science Practices Factor Loadings		
Assessment modality factors	Identifying	Using	Conducting
Static	0.702	0.793	0.702
Active	0.654	0.766	0.739
Interactive	0.664	0.758	0.673

Table 6 shows the multitrait-multimethod (MTMM) correlation matrix. The upper table in Table 4 focuses on how the intended constructs (practices) hold using different methods (modalities). The figures in bold represent the within-trait, cross-method correlations and the values (which range from .53 to .69) show moderate positive correlations, which provides evidence of convergent validity.

Table 6. Multitrait-multimethod correlation matrix for three constructs measured by three tasks

		Static			Active			Interactive			
		Identify	Use	Conduct	Identify	Use	Conduct	Identify	Use	Conduct	Identify
Static	Identify	1.00									
	Use	0.56	1.00								
	Conduct	0.53	0.56	1.00							
Active	Identify	**0.62**	0.00	0.06	1.00						
	Use	-0.01	**0.54**	0.02	0.49	1.00					
	Conduct	0.04	0.00	**0.60**	0.54	0.59	1.00				
Interacti	Identify	**0.60**	0.00	0.06	**0.69**	-0.01	0.05		1.00		
	Use	-0.01	**0.53**	0.03	-0.02	**0.53**	0.03		0.49	1.00	
	Conduct	0.05	0.00	**0.59**	0.08	0.03	**0.62**		0.52	0.55	1.00

		Identify			Use			Conduct		
		Static	Active	Interactive	Static	Active	Interactive	Static	Active	Interactive
Identify	Static	1.00								
	Active	0.62	1.00							
	Interactive	0.60	0.69	1.00						
Use	Static	**0.56**	0.00	0.00	1.00					
	Active	-0.01	**0.49**	-0.01	0.54	1.00				
	Interactive	-0.01	-0.01	**0.49**	0.53	0.53	1.00			

Conduct	Static	**0.53**	0.06	0.05	**0.56**	0.02	0.03	1.00		
	Active	0.04	**0.54**	0.05	0.00	**0.59**	0.03	0.60	1.00	
	Interactive	0.05	0.07	**0.51**	0.00	0.03	**0.55**	0.59	0.62	1.00

The lower table in Table 6 provides information about how each science practice construct is tied to each assessment modality. The figures that are in bold within the shaded areas represent the within-method, cross-trait correlations. The correlation between the science practice constructs of *Identifying* and *Using* is lower for the interactive modality (.53) than for the static and active modality (both are .56). This is the same for the relationship between *Identifying* and *Conducting* (shaded in light orange and in bold) as well as for the relationship between Using and Conducting (shaded in light purple and in bold). Overall, the lowest correlations between the science practice constructs occur in the interactive assessment modality (.49 for *Identify/Use*, .51 for *Identify/Conduct* and .55 for *Use/Conduct)*, although the correlation for *Identify/Use* in the active mode (.49) was as low as in the interactive modality.

Multidimensional IRT

Another way to compare how well the three assessment modalities are measuring the individual science practice constructs is to examine the overall reliability for the 24 items that made up each assessment modality. Tables 7a - 7c show the inter-correlations between the three science practice constructs under each of the three assessment modalities.

Table 7a. Construct Inter-correlations for Static Modality

	Identifying	Using	Conducting
Identifying	1.00	.92	**.82**
Using		1.00	.92
Conducting			1.00

Table 7b. Construct Inter-correlations for Active Modality

	Identifying	Using	Conducting
Identifying	1.00	**.79**	**.79**
Using		1.00	.97
Conducting			1.00

Table 7c. Construct Inter-correlations for Interactive Modality

	Identifying	Using	Conducting
Identifying	1.00	.83	**.72**
Using		1.00	.86
Conducting			1.00

The overall Cronbach's alpha reliability was highest for the 24 items in the interactive assessment mode (.85) compared to .82 for the static mode and .81 for the active mode. Table 8 shows another type of reliability coefficient estimated by the ConQuest IRT analysis software,

the Expected A-Posteriori (EAP)/Plausible Values (PV). This is derived from estimates of the conditioned posterior ability distributions for each person on the multiple dimensions (see Adams, Wu, & Macaskill, 1997). From the conditioned student posterior ability distributions, plausible values (PVs) are randomly drawn. The mean value of the conditioned posterior distribution is known as the Expected A-Posteriori (EAP) estimate. The EAP/PV reliability coefficient represents how well the persons (students) are separated by the measures of each of the science practice constructs. Table 8 shows the EAP/PV coefficients for each assessment modality. For each science practice construct (represented in the columns of the table), the highest reliability is shown in bold. For *Identifying* and *Using* practice constructs, the static modality of assessment was the most reliable (.76 for *Identifying* and .79 for *Using*), but for the *Conducting* construct, the interactive assessment modality was most reliable (.82). These findings converge with those from the G-study and the confirmatory factor analyses confirming that the interactive modality most distinctly measured the Conducting Inquiry construct.

Table 8. EAP/PV Reliability Coefficients for Static, Active and Interactive Modalities

	Identifying	Using	Conducting
Static	.76	.79	.77
Active	.74	.76	.77
Interactive	.74	.77	.82

Discussion and Implications

With the increasing interest in the use of technology to create assessments that measure skills that are hard to assess in traditional static modalities, this study suggests that engaging students in interactive assessments may provide a better estimate of their more complex inquiry skills than active or static formats do. We also know from our review of extant items that such interactive modalities are currently not widely used in science assessment, if at all. The study contributes a synthesis of research-based principles and literature sources for informing the design of next generation assessments, an analysis of the extent to which current large-scale science tests address integrated knowledge and inquiry practices, and an empirical study of the affordances of dynamic and interactive modalities for measuring distinct science practice constructs. These outcomes of the *Foundations of 21st Century Science Assessments* project provide guidelines for designing the next generation of science assessments, information suggesting that extant tests do not cover systems thinking or the practices of science, and evidence supporting claims that the affordances of dynamic, interactive, complex assessment tasks can improve the measurement of science inquiry practices.

We note that the design principles presented in this article relate to the types of summative assessment in the three modalities compared in this study. Literature on the affordances of dynamic, interactive modalities for formative and adaptive purposes was synthesized by the project and will be reported elsewhere. Relatively little research has studied the interaction of multiple media such as text, graphics, static and dynamic perceptual cueing in complex tasks. There is considerable research to be done on the functions of multiple representations and interactive interfaces in learning and assessments of science systems and practices (Buckley, in press).

The sample of extant test items analyzed was limited to the middle school level and the two topics of ecosystems and chemistry. Further analyses would need to determine if the uneven coverage of some of the core ideas (e.g., system models) and science practices (e.g. designing, conducting, and critiquing) in the new *Framework for K-12 Science Education* will be found in a wider sample of items. Few extant items were found in the active modality and none were accessible in the interactive modality. The analysis found that items combining text and graphics did not exemplify best use of the design principles. As more innovative science assessments include dynamic, interactive modalities, these new item formats can be examined to determine their utilization of research-based design principles.

This study provides rare large-scale evidence that interactive assessments may be more effective than static assessments at discriminating student proficiencies across different types of science practices. Studies comparing item formats have primarily been within the static modality (selected vs. constructed responses) or between performance assessments and conventional tests. This study extends the comparison of task and item design to complex tasks, system models, inquiry practice constructs, and the dynamic and interactive affordances of technology-based complex science assessment tasks. The three modality versions compared (static, active, and interactive) were carefully constructed to keep the representations of the science ecosystem parallel. Thus, all three versions depicted the ecosystems with parallel stylistic images of the organisms, tables, graphs, and screen layouts. In contrast, "found" ecosystem items varied in the representation of the ecosystem, for example, by presenting a food web as a set of boxes, text organism names, or pictures of organisms. This study aimed for comparable representations of the ecosystems so that the variables would be the extent of learner control (static, active, interactive) and the dynamic level of the ecosystem presentation, i.e. a still image, an animation, or a dynamic, changing display. Research on design variations within next generation assessments will face similar methodological challenges.

The study of alternative complex task and item formats also presents analysis challenges. In this study, a combination of methods—a generalizability study, IRT, and confirmatory factor analyses—examined the measurement properties of the modalities through different lenses. Examining the convergent, discriminate, and construct validity of complex, dynamic assessments poses challenges for the measurement community.

Conclusions

This project integrated research on learning in rich multimedia environments with rigorous analyses of extant static and dynamic science assessment tasks, then used evidence-centered assessment design methods as the framework for developing and establishing the technical quality of reusable task designs for assessing complex science learning. We believe this will make a significant contribution to the field by moving the state of technology-based item development from art to principled practice through identifying relevant findings from research on model-based reasoning and multimedia learning that affect the design of assessments of learning, retrieval, and transfer.

Within a limited sample of items, current science tests did not seem to address some of the valued knowledge and practices called for in the new *Framework for K-12 Science Education*. Therefore, the next generation of science assessments will need to address both a broader range of standards and innovative methods for assessing them.

The study provides much needed empirical evidence of the affordances of dynamic and interactive assessments for discriminating among science knowledge and inquiry skills. The

results suggest that static assessments are not as effective as active and interactive assessments for differentiating between factual knowledge and the ability to apply that knowledge in meaningful contexts. Our study found that the interactive task sets that serve as a basis of the interactive assessments were more effective than either static or active assessments at uniquely measuring students' ability to engage in inquiry practices. Therefore, assessment developers who wish to design assessments of science inquiry skills should consider the use of active and interactive assessment tasks.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. DRL-0814776 awarded to WestEd, Edys Quellmalz, Principal Investigator and co-Principal Investigators Michael Timms, Jodi Davenport, and George DeBoer. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Adams, W. K., Reid, S., Lemaster, R., McKagan, S. B., Perkins, K. K., Dubson, M., et al. (2008). A study of educational simulations part 1—engagement and learning. *Journal of Interactive Learning Research, 19*(3), 397-419.

Adams, R. J., Wu, M. L., & Macaskill, G. (1997). Scaling methodology and procedures for the mathematics and science scales. In M. O. Martin & D. L. Kelly (Eds.), TIMSS technical report volume II: Implementation and analysis (pp. 111-145). Chestnut Hill, MA: Boston College.

Ainsworth, S. (2008). The educational value of multiple-representations when learning complex scientific concepts. In J. K. Gilbert, M. Reiner & M. Nakhleh (Eds.), *Visualization: Theory and practice in science education*. New York: Springer.

American Association for the Advancement of Science (AAAS). (1993). Benchmarks for science literacy. New York, NY: Oxford University Press.

Bangert-Downs, R. L., Kulik, C. L. C, Kulik, J. A., & Morgan, M. T. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research, 61*(2), 213–238.

Bell, P., & Linn, M. C. (2000). Beliefs about science: How does science instruction contribute? In B. K. Hofer & P. R. Pintrich (Eds.), *Personal epistemology: The psychology of beliefs about knowledge and knowing*. Mahwah, NJ: Lawrence Erlbaum Associates.

Betrancourt, M. (2005). The animation and interactivity principles. In R. E. Mayer (Ed.), *Handbook on multimedia learning*. New York: Cambridge University Press.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (2000). *How people learn: Brain, mind, experience, and school*. Washington, DC: National Academy Press.

Brennan, R. (2001). mGENOVA. Retrieved from http://www.education.uiowa.edu/centers/casma/computer-programs.aspx.

Buckley, B. C. (in press). Supporting and Assessing Complex Biology Learning with Computer-based Simulations and Representations. In D. Treagust & C.-Y. Tsui (Eds.), *Multiple Representations in Biological Education*: Springer.

Buckley, B. C. (in press, 2011). Model-based Learning. In N. Seel (Ed.), *Encyclopedia of the Sciences of Learning*: Springer Science.

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), 166 - 190.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*. 56:81-105.

Clark, R. C., & Mayer, R. E. (2011). *E-learning and the science of instruction : Proven guidelines for consumers and designers of multimedia learning*. San Francisco, CA: Pfeiffer.

Clement, J. (1989). Learning via model construction and criticism: Protocol evidence on sources of creativity in science. In J. A. Glover, R. R. Ronning & C. R. Reynolds (Eds.), *Handbook of creativity: Assessment, theory and research* (pp. 341–381). New York: Plenum Press.

Clement, J. J., & Rea-Ramirez, M. A. (Eds.). (2008). *Model Based Learning and Instruction in Science*. London: Springer.

College Board. (2009). *Science: College Boards standards for college success.* Retrieved from http://professionals.collegeboard.com/profdownload/cbscs-science-standards-2009.pdf.

Collins, A., Brown, J. S., & Newman, S. (1989). Cognitive apprenticeship: Teaching the craft of reading, writing, and mathematics. In L. B. Resnick (Ed.), *Knowing, learning,a nd instruction: Essays in honor of Rober Glaser* (pp. 453–494). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Dassa, C., Vazquez-Abad, J., & Ajar, D. (1993). Formative assessment in a classroom setting: From practice to computer innovations. *Alberta Journal of Educational Research*, 111.

Darling-Hammond, L. (2010). *Performance counts: Assessment systems that support high-quality learning.* CCSSO: Washington, DC.

Darling-Hammond, L., & Pecheone, R. (2010). *Developing an internationally comparable balanced assessment system that supports high-quality learning.* Presented at the National Conference on Next Generation K–12 Assessment Systems, Center for K–12 Assessment & Performance Management with the Education Commission of the States (ECS) and the Council of Great City Schools (CGCS), Washington, DC.

Dede, C. (2009). Immersive interfaces for engagement and learning. *Science, 323*(5910), 66–69.

Donovan, M. S., & Bransford, J. D., (2005). *How students learn: Science in the classroom.* Washington, DC: The National Academies Press.

Duschl, R. A., & Osborne, J. (2002). Supporting and promoting argumentation discourse in science education. *Studies in Science Education, 38*, 39–72.

Duschl, R. A., Schweingruber, H. A., & Shouse, A. W. (2007). *Taking science to school: Learning and teaching science in grades K–8*. Washington, DC: The National Academies Press.

Edelson, D., & Reiser, B. J. (2006). Making authentic practices accessible to learners: Design challenges and strategies. In K. Sawyer (Ed.), *Cambridge handbook of the learning sciences* (pp. 335–354). New York: Cambridge University Press.

Geier, R., Blumenfeld, P., Marx, R., Krajcik, J., Fishman, B., & Soloway, E. (2008). Standardized test outcomes of urban students participating in standards and project based science curricula. *Journal of Research in Science Teaching, 45*(8), 922–939.

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.

Hegarty, M. (2004). Dynamic visualizations and learning: Getting to the difficult questions. *Learning & Instruction, 14*(3), 343–351.

Hmelo-Silver, C. E., Jordan, R., Liu, L., Gray, S., Demeter, M., Rugaber, S., et al. (2008). Focusing on function: Thinking below the surface of complex science systems. *Science Scope, 31*(9), 27–35.

Horwitz, P., Gobert, J. D., Buckley, B. C., & O'Dwyer, L. M. (2010). Learning genetics with dragons: From computer-based manipulatives to hypermodels. In M. J. Jacobson & P. Reimann (Eds.), *Designs for learning environments of the future: International perspectives from the learning sciences* (pp. 61–87). New York: Springer.

Horwitz, P., Gobert, J., Buckley, B. C., & Wilensky, U. (2007). *Modeling across the curriculum annual report to NSF*: The Concord Consortium.

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.

Jacobson, M. J. (2001). Problem solving, cognition, and complex systems: Differences between experts and novices. *Complexity, 6*(3), 41–49.

King, K. (2011). *Balanced, multilevel science assessment systems.* Presented at the National Conference on Student Assessment. Orlando, FL.

Kolodner, J. L., Camp, P. J., Crismond, D., Fasse, B., Gray, J., Holbrook, J., et al. (2003). Problem-based learning meets case-based reasoning in the middle-school science classroom: Putting learning by design into practice. *The Journal of the Learning Sciences, 12*(4), 495–547.

Koomen, M. (2006). *The development and implementation of a computer-based assessment of science literacy in PISA 2006*. Paper presented at the Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Krajcik, J., Blumenfeld, P. C., Marx, R. W., Bass, K. M., & Fredricks, J. (1998). Inquiry in project-based science classrooms: Initial attempts by middle school students. *Journal of the Learning Sciences, 7*(3&4), 313–350.

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.

Kriz, S. & Hegarty, M. (2004). Constructing and revising mental models of a mechanical system: The role of domain knowledge in understanding external visualizations. In K. Forbus, D. Gentner & T Regier (Eds.) *Proceedings of the 26th Annual Conference of the Cognitive Science Society.* Mahwah, NJ: Lawrence Erlbaum Associates.

Kühl, T., Scheiter, K., Gerjets, P., & Edelmann, J. (2011). The influence of text modality on learning with static and dynamic visualizations. *Computers in Human Behavior, 27*, 29-35.

Larkin, J., & Simon, H. (1987). Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science, 11*(1), 65–100.

Lee, O. (2002). Science inquiry for elementary students from diverse backgrounds. In W. G. Secada (Ed.), *Review of research in education* (Vol. 26, pp. 23–69). Washington, DC: American Educational Research Association.

Lee, H., Plass, J. L., & Homer, B. D. (2006). Optimizing cognitive load for learning from computer-based science simulations. *Journal of Educational Psychology, 98*(4), 902-913.

Lehrer, R., & Schauble, L. (2002). Symbolic communication in mathematics and science: Co-constituting inscription and thought. In E. D. A. J. Byrnes (Ed.), *Language, literacy, and cognitive development. The development and consequences of symbolic communication* (pp. 167–192). Mahwah, NJ: Lawrence Erlbaum Associates.

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.

Li, M., & Shavelson, R. J. (2001). *Examining the linkage between sceince acheivement and assessment.* Paper presented at the American Educational Research Association, Seattle, WA.

Linn, M. C., Bell, B., & Davis, E. A. (2004). *Internet environments for science education.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Linn, M.C., & Eylon, B.-S. (2011). *Science learning and instruction: Taking advantage of technology to promote knowledge integration*. New York: Routledge.

Liu, O. L., Lee, H.-S., & Linn, M. C. Measuring knowledge integration: Validation of four-year assessments. *Journal of Research in Science Teaching, 48*(9), 1079–1107.

Loehlin, J. C. (1998). *Latent variable model: An introduction to factor, path, and structural analysis*. Mahwah, N.J.: Lawrence Erlbaum.

Lowe, R., & Schnotz, W. (2008). *Learning with animation: Research implications for design*. Cambridge; New York: Cambridge University Press.

Lowe, R. K. & Schnotz, W. (Eds) (2007) *Learning with animation*. New York: Cambridge University Press.

Mayer, R. E. (Ed.) (2005). *The Cambridge handbook of multimedia learning*. New York: Cambridge University Press.

Mayer, R. E., & Johnson, C. (2008). Revising the redundancy principle in multimedia learning. *Journal of Educational Psychology, 100*, 380–386.

Marx, R. W., Blumenfeld, P. C., Krajcik, J. S., Fishman, B., Soloway, E., Geier, R., et al. (2004). Inquiry-based science in the middle grades: Assessment of learning in urban systemic reform. *Journal of Research in Science Teaching, 41*(10), 1063–1080.

Messick, S. (1994) The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 12–23.

Metz, K. E. (2004). Children's understanding of scientific inquiry: Their conceptualization of uncertainty in investigations of their own design. *Cognition and Instruction, 22*(2), 219–290.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). *A brief introduction to evidence-centered design*: Educational Testing Service.

National Assessment Governing Board (NAGB). (2008). *Science framework for the 2009 national assessment of educational progress*. Washington DC: National Assessment Governing Board.

National Research Council (NRC). (2011). *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas*. Washington, DC: The National Academies Press.

Nersessian, N. J. (2008). *Creating scientific concepts*. Cambridge, MA: MIT Press.

Pashler, H., Bain, P., Bottge, B., Graesser, A., Koedinger, K., McDaniel, M., et al. (2007). *Organizing instruction and study to improve student learning* (No. NCER 2007-2004). Washington, DC: National Center for Education Research, Institute of Education Sciences, U.S. Department of Education.

Pedone, R., Hummel, J. E., & Holyoak, K. J. (2001). The use of diagrams in analogical problem solving. *Memory and Cognition, 29*, 214–221.

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., 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*. Unpublished manuscript, Washington DC.

Quellmalz, E. S., & Pellegrino, J. W. (2009). Technology and testing. *Science, 323*, 75–79.

Quellmalz, E., Schank, P., Hinojosa, T., & Padilla, C. (1999). *Performance assessment links in science (PALS)* (No. ERIC Digest Series E-DO-TM-99-04). College Park, MD: ERIC Clearinghouse on Assessment and Evaluation.

Quellmalz, E. S., Timms, M. J., & Buckley, B. (2010). The promise of simulation-based science assessment: The calipers project. *International Journal of Learning Technology, 5*(3), 243–263.

Quellmalz, E. S., Timms, M. J., Buckley, B. C., Davenport, J., Loveland, M., & Silberglitt, M. D. (2011). 21st Century Dynamic Assessment. In J. Clarke-Midura, M. Mayrath & C. Dede (Eds.), *Technology-Based Assessments for 21st Century Skills: Theoretical and Practical Implications from Modern Research* (pp. 55–89): Information Age.

Quellmalz, E. S., Timms, M. J., Silberglitt, M. D., & Buckley, B. C. (in press). Science assessments for all: Integrating science simulations into balanced state science assessment systems. *Journal of Research in Science Teaching*.

Quintana, C., Reiser, B. J., Davis, E. A., Krajcik, J., Fretz, E., Duncan, R. G., et al. (2004). A scaffolding design framework for software to support science inquiry. *Journal of the Learning Sciences, 13*(3), 337–386.

Reif, F. (2008). *Applying cognitive science to education: Thinking and learning in scientific or other domains*. Cambridge, Mass.; London: MIT.

Reiser, B. J. (2004). Scaffolding complex learning: The mechanisms of structuring and problematizing student work. *Journal of the Learning Sciences, 13*(3), 273–304.

Rivet, A., & Krajcik, J. S. (2004). Achieving standards in urban systemic reform: An example of a sixth grade project-based science curriculum. *Journal of Research in Science Teaching, 41*(7), 669–692.

Schwartz, D. L., & Black, T. (1999). Inferences through imagined actions: Knowing by simulated doing. *Journal of Experimental Psychology. Learning, Memory & Cognition, 25*(1), 116.

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.

Shavelson, R.J. (2006). On the Integration of Formative Assessment in Teaching and Learning: Implications for New Pathways in Teacher Education. In F. Oser, F. Achtenhagen, and U. Renold, eds., *Competence-Oriented Teacher Training: Old Research Demands and New Pathways*. Utrecht, The Netherlands: Sense Publishers.

Shavelson, R. J., Ruiz-Primo, M. A., & Wiley, E. W. (2005). Windows into the mind. *Higher Education, 49*(4), 413–430.

Slotta, J. D., & Chi, M. T. H. (2006). Helping students understand challenging topics in science through ontology training. *Cognition and Instruction, 24*(2), 261–289.

Simon, H. A. (1980). Problem solving and education. In D.T. Tuma and F. Reif (Eds.), *Problem solving and education: Issues in teaching and research* (pp. 81–96). Hillsdale, NJ: Erlbaum.

Smith, C. L., Wiser, M., Anderson, C. W., & Krajcik, J. (2006). Implications of research on children's learning for standards and assessment: A proposed learning progression for matter and the atomic-molecular theory. *Measurement: Interdisciplinary Research and Perspectives, 4*(1–2), 1–98.

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.

Tabak, I., & Reiser, B. J. (2008). Software-realized inquiry support for cultivating a disciplinary stance. *Pragmatics and Cognition, 16*(2), 307–355.

Tversky, B., Heiser, J., Lozano, S., MacKenzie, R., & Morrison, J. (2008). Enriching animations. In R. K. Lowe & W. Schnotz (Eds.), *Learning with animation: Research and design implications*. New York: Cambridge University Press

Van Merriënboer, J.J.G., & Kester, L. (2005). The four-component instructional design model: Multimedia principles in environments for complex learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 71–93). New York: Cambridge University Press.

Vattam, S. S., Goel, A. K., Rugaber, S., Hmelo-Silver, C. E., Jordan, R., Gray, S., et al. (2011). Understanding complex natural systems by articulating structure-behavior- function models. *Journal of Educational Technology & Society, 14*(1), 66–81.