Assessment, Technology, and Change

Jody Clarke-Midura and Chris Dede
Harvard Graduate School of Education

Abstract

Despite three decades of advances in information and communications technology (ICT) and a generation of research on cognition and new pedagogical strategies, the field of assessment has not progressed much beyond paper-and-pencil item-based tests. Research has shown these instruments are not valid measures of sophisticated intellectual performances. Simply using technology to deliver automated versions of item-based tests does not realize the full power of ICT to innovate in assessment via providing rich experiences that enable observing and analyzing student performances. To illustrate this approach, we describe our early research on using immersive technologies to develop virtual performance assessments. (Keywords: Assessment, technology, virtual worlds, science inquiry)

Despite almost three decades of advances in information and communications technology (ICT) and a generation of research on cognition and on new pedagogical strategies, the field of assessment has not progressed much beyond paper-and-pencil item-based tests whose fundamental model was developed a century ago. In 2001, the National Research Council (NRC) published a report, Knowing What Students Know, that highlighted current innovative projects using technology to assess learning and foreshadowed how further advances in technology and statistical analysis could provide new models for assessment. However, not until recently did state, national, and international high-stakes testing programs start to deliver assessments via technology. For example, in 2006 the Programme for International Student Assessment (PISA) piloted online versions of its items preparatory to moving into online delivery. In the United States, the National Assessment of Educational Progress (NAEP) recently piloted technology-based items in math and literacy, and developers are currently designing technology-based items for science. Also, states such as Minnesota and North Carolina are starting to use technology-based items in accountability settings (Quellmalz & Pellegrino, 2009). However, using technology to deliver automated versions of item-based paper-and-pencil tests does not realize the full power of information and communication technologies (ICT) to innovate via providing richer observations of student learning.

This paper describes research underway that is attempting a breakthrough in the use of technology to improve assessment dramatically beyond the century-old methods in widespread use today. Such an advance in assessment

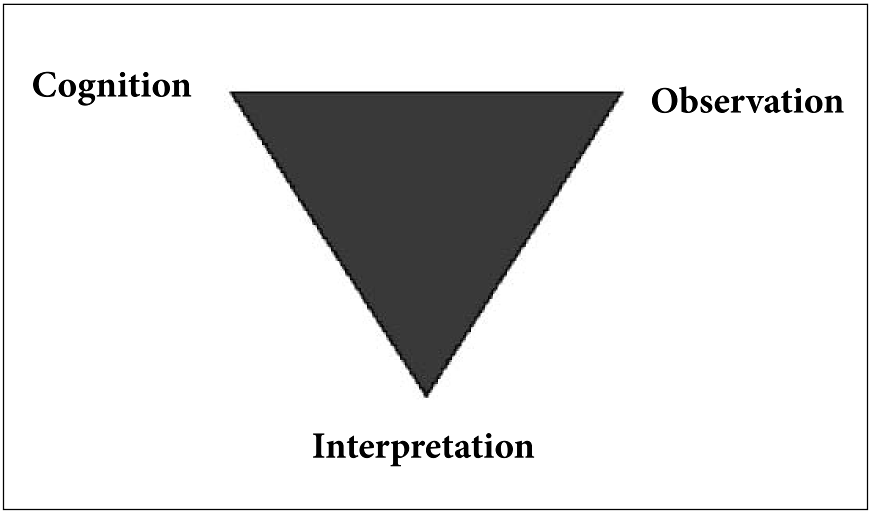


Figure 1. Assessment Triangle.

is greatly needed because current methods of testing are incapable of validly measuring sophisticated intellectual and psychosocial performances. For example, studies of national tests, such as the NAEP, showed the items related to scientific inquiry did not align with the inquiry content they were supposed to be measuring (Haertel, Lash, Javitz, & Quellmalz 2006; Quellmalz & Haertel, 2004; Quellmalz, Kreikemeier, DeBarger, & Haertel, 2006). These studies recommended that test designers redesign some of their items and integrate technology-based measures. But why are item-based, paper-and-pencil tests inadequate for important student outcomes, such as scientific inquiry and 21st-century skills?

The NRC report, Knowing What Students Know, depicted the "Assessment Triangle" (NRC, 2001), which identified three key components of assessment (see Figure 1).

We cannot directly inspect what students know or do not know. Like Sherlock Holmes solving mysteries, assessment involves indirect reasoning from evidence—developing a model of cognition reflecting the knowledge a learner is to master, collecting observations of a student's statements and behaviors, and interpreting the extent to which those statements and behaviors match the expert model.

Over the last few decades, cognitive scientists have greatly increased the depth and validity of their models, and psychometricians have made major advances in interpretive analytics. However, the observation part of the triangle, centered on paper-and-pencil item-based tests, has remained weak for about a century. These tests cannot generate a rich range of observations; students' forced choice among a few predetermined options is a weak observation of whether they have mastered a sophisticated skill involving advanced knowledge. Without detailed observations that document every

Copyright © 2010 ISTE (International Society for Technology in Education), 800.336.5191 (U.S. & Canada) or 541.302.3777 (Int'l), iste@iste.org, www.iste.org. All rights reserved.

aspect of a learner's performance, little is available to compare to the highly specified cognitive model using advanced interpretive analytics. Attempts to improve assessment have repeatedly foundered on this problem of impoverished observations of student performances within the rigorous conditions required to ensure the fairness, reliability, and validity of sophisticated intellectual assessments. Our research is attempting a breakthrough in assessment because technical capabilities now exist for providing rich observations about student learning.

When it comes to testing in an accountability setting, multiple-choice tests have been the favored choice because they have satisfied psychometric criteria, are more cost effective, and are easier to scale. Movements for more authentic or performance-based assessments that are better aligned with how students learn rarely get enough traction against their multiple-choice counterparts. For example, as discussed in detail below, performance-based measures using physical objects in real-world settings were shown to be not as psychometrically reliable or practical as item-based tests, are expensive, and are burdened with task-dependency (Linn, 2000; Shavelson, Ruiz-Primo, & Wiley, 1999; Stecher & Klein, 1997).

However, advances in technology and statistics are creating new possibilities and promises for assessment. The type of observations and evidence of learning that technology-based assessments allow is unparalleled. For example, research on immersive environments and mediated experiences proves that one can create environments capable of capturing observations and studying authentic behaviors not possible in a conventional classroom setting (Clarke, 2009b; Ketelhut, Dede, Clarke, Nelson, & Bowman, 2008). Based on the immersive interface that underlies virtual worlds such as Second Life and World of Warcraft, virtual environments allow the enactment of complex situations with tacit clues, simulation of scientific instruments, virtual experimentation, simulated collaboration in a team, and adaptive responses to students' choice—all captured and recorded in data streams (Dede, 2009). Current technological advances offer exciting opportunities to design assessments that are active and situative, and that measure complex student knowledge and provide rich observations for student learning.

In this article, we present a model for how technology can provide more observations about student learning than current assessments. To illustrate this approach, we describe our early research on using immersive technologies to develop virtual performance assessments. We are using the Evidence Centered Design (ECD) framework (Mislevy, Steinberg, & Almond, 2003) to develop interactive performance assessments for measuring scientific inquiry that are more reflective of the situative, complex performances that scientists and scholars expert in inquiry learning call for students to master. In the following sections, we describe the background and context of our work, and then depict how immersive technologies and mediated performances may improve assessment. We conclude with suggestions for future research.

The Inadequacy of Conventional Approaches to Assessment in 21st-Century Education

Paper-and-pencil tests are barely adequate to measure the minimum competencies required for low-level roles in industrial settings and fall woefully short of providing measures of the sophisticated knowledge and skills students need for 21st-century work and citizenship. States' high-stakes psychometric tests are typically based on multiple-choice and short-answer items that have no mechanism for assessing attainment of higher-order understandings and performances (National Research Council, 2001). As a result, the curriculum is crowded with low-level facts and recipe-like procedures (e.g., In what year did Columbus discover America? What are the seven steps of historical inquiry?), as opposed to nuanced understandings and performances (i.e., What confluence of technological, economic, and political forces led to the age of exploration around the end of the 15th century? By what process of interpreting of historical data did you reach this conclusion?). Even the essay section of high-stakes tests emphasizes simple execution based on recipe-like formulas for each paragraph, rather than allowing students to demonstrate sophisticated forms of rhetorical prowess.

State curriculum standards in each discipline are typically neither interrelated nor prioritized to emphasize core understandings and performances all students will need to succeed in the 21st century (Aspen Institute, 2007). Although professional organizations such as the American Association for the Advancement of Science or the National Council for the Teaching of Mathematics make attempts at integration and standardization in national standards, in practice this level of understandings and performances is ignored in classroom teaching because the high-stakes tests provide no vehicle for measuring student progress on them.

Because of the accountability systems linked to students' performance on these high-stakes tests, teachers are using weak but rapid instructional methods, such as lecture and drill-and-practice, to race through the glut of recipes, facts, and test-taking skills they are expected to cover. Despite research indicating that guided inquiry, collaborative learning, mentoring, and apprenticeships are far more effective pedagogical strategies, introducing these into school settings is difficult given the crowded curriculum and the need to prepare students for high-stakes tests. Simply delivering required information for students' passive absorption takes every second of instructional time. Teachers have no means by which to prioritize what understandings and performances to emphasize in terms of 21st-century citizenship; workplace capabilities for the global, knowledge-based economy; and lifelong learning (Dede, 2007, in press), and they do not assess students' abilities to transfer their skills to real-world situations.

These summative, "drive-by" tests provide no diagnostic, just-in-time feedback that could help teachers aid struggling students. In addition, while some paper and pencil assessments, such as the Programme for International

Student Assessment (PISA) test, emphasize core ideas and measure at least a few higher-order thinking skills, many state legislatures have allocated such limited resources for test development that the resulting instruments often measure only a random assortment of low-level skills and content (Nicols, Glass, & Berliner, 2005). Furthermore, policies such as financial incentives for teachers and districts to raise test scores can exacerbate already troubling differences in educational outcomes, promoting the abandonment of the very at-risk students the NCLB legislation was intended to help (Confrey & Maker, 2005).

Even though modern interactive media could aid with these shortfall of current high-stakes tests, the use of ICT applications and representations is generally banned from testing. Rather than measuring students' capacities to use tools, applications, and media effectively, various forms of mediated interaction are typically not assessed. In other words, the effects from technology usage (what one can accomplish without tools) are measured, but the effects with technologies essential to effective practice of a skill are not (Salomon, 1993).

Historical Perspective on the Value and Challenges of Performance Assessments

Research has documented that higher-order thinking skills related to sophisticated cognition (e.g., inquiry processes, formulating scientific explanations, communicating scientific understanding, approaches to novel situations) are difficult to measure with multiple-choice or even constructed-response paper-and-pencil tests (NRC, 2006; Quellmalz & Haertel, 2004; Resnick & Resnick, 1992). These tests also demonstrate limited sensitivity to discrepancies between inquiry- and non-inquiry-based science instruction (Haertel, Lash, Javitz, & Quellmalz, 2006). In the 1990s, there was a movement toward developing alternate assessments in science education that measured students' conceptual understanding and higher-level skills, such as problem solving (Linn, 1994). Numerous studies assessed the reliability and construct validity of these performance assessments, as well as the feasibility (i.e., cost effectiveness and practicality) of using them on a large scale (Linn, 2000). Although research supports performance tasks as valuable both for aiding learning and for providing formative, diagnostic feedback to teachers about ongoing student attainment, when used as summative assessments, performance tasks were found to be not as cost-effective as multiple-choice tests (Stecher & Klein, 1997). Also, performance assessments had troubling issues around task sampling variability (Shavelson, Baxtor, & Gao, 1993). These studies found that students' outcomes on performance tasks varied substantially from one task to another. Ideally, one would want students to perform the same on all tasks. Another problem with performance assessments, also related to sampling variability, was occasion-sampling variability (Cronbach, Linn, Brennan, & Haertel, 1997). Studies found that student performances varied on one testing occasion to another. Shavelson, Ruiz-Primo, and Wiley (1999) found that task sampling and occasion sampling were confounded. Thus, it was difficult to distinguish if the error was due to testing occasion or sampling variability.

As one illustration, Shavelson and colleagues conducted a series of studies in the 1990s in which they compared computer-simulated performance assessments to paper-based performance assessments (Baxter, 1995; Baxter & Shavelson, 1994; Pine, Baxter, & Shavelson, 1993; Rosenquist, Shavelson, & Ruiz-Primo, 2000; Shavelson, Baxter, & Pine, 1991). Their findings suggested that hands-on and virtual investigations were not tapping the same knowledge as paper-based assessments (Shavelson, Baxter, & Pine, 1991), that prior knowledge and experience influence how students solve the problem (Shavelson, Baxter, & Pine, 1991), and that the volatility of student performance limits the exchangeability of any methods used for delivering tasks (direct observation, notebook, computer simulation, paper-and-pencil methods, etc.) (Shavelson et al., 1999).

As a result, they suggest that multiple assessments are needed to make adequate observations of student performance. Different methods for delivering tasks may allow for a variety of evidence that can be used to triangulate student performance. As we will discuss later, virtual performance assessments developed in immersive technologies are able to provide a variety of evidence and methods under one assessment. Advances in technology are allowing us to create the types of assessments Shavelson and his colleagues envisioned.

In summary, the goal of an assessment is to provide valid inferences related to particular expectations for students (Linn et al., 2002). While an assessment can serve multiple purposes, it is not possible for one assessment to meet all purposes; for example, an assessment providing information about students' deep conceptual understanding that can be used by educators to guide instruction would be different from an assessment that provides an evaluation of an educational system for policymakers (NRC, 2001). To meet the requirements of No Child Left Behind (NCLB), it is recommended that states develop a variety of assessment strategies, including performance assessments that collectively will fulfill requirements (NRC, 2006). However, there are substantial challenges of practicality, cost, and technical quality involved in achieving this goal using various types of conventional measures. Assessments based on immersive technologies and mediated performances are potentially more practical, cost effective, valid, and reliable than performance assessments that were developed and studied in the past. We explain why below.

Research Initiatives to Improve Assessments

The most prevalent current use of technology in large-scale testing involves support for assessment logistics related to online delivery, automated scoring, and accessibility. For example, Russell and colleagues are building accessibility for all students into the design of the items via technology (http://nimbletools.com). The technology is flexible enough that it allows customization to be built in to the construction of the item. Such customizations include: audio text or text-to-speech, magnification, masking tools that allow students to focus on specific parts of the test item, presentation settings (color overlays, contrast, etc.), sign language, and auditory calming. This all happens at the design level of the item, creating a cost-effective way to customize items for a wide variety of students.

However, research on next-generation assessments is breaking the mold of traditional testing practices. Innovative assessment formats, such as simulations, are being designed to measure complex knowledge and inquiry previously impossible to test in paper-based or hands-on formats. These new assessments aim to align summative assessment more directly to the processes and contexts of learning and instruction (Quellmalz & Pellegrino, 2009).

As Quellmalz, Timms, and Schneider discuss (2009, p. 7):

> A number of large-scale testing programs have begun to design innovative problem sets and item types that promise to transform traditional testing. The area of science assessment is pioneering the exploration of innovative problem types and assessment approaches across K–12. In 2006, the Programme for International Student Assessment (PISA) pilot tested the Computer-Based Assessment of Science (CBAS) with the aim of testing science knowledge and inquiry processes not assessed in the PISA paper-based test booklets. CBAS tasks included scenario based item and task sets such as investigations of the temperature and pressure settings for a simulated nuclear reactor. The 2009 National Assessment of Educational Progress (NAEP) Science Framework and Specifications proposed designs for Interactive Computer Tasks (ICT) to test students' ability to engage in science inquiry practices. These innovative formats were included in the 2009 NAEP science administration.

A few years ago, NAEP published their framework for establishing a new science assessment in 2009 that calls for multiple modes of assessment, including interactive computer assessments (National Assessment Governing Board, 2004, 2007). The report cites four reasons for rethinking the assessment framework: publication of national standards for science literacy since the previous framework, advances in both science and cognitive research, growth in national and international science assessments, and increases in innovative assessment approaches.

The NAEP report also states the need for studies that compare hands-on assessments to computer-based assessments. In addition to citing the importance of assessments related to inquiry, the experts involved suggest that test-makers introduce interactive computer tasks to assess students' knowledge, skills, and abilities related to the following situations (National Assessment Governing Board, 2007, p. 107):

- For scientific phenomena that cannot easily be observed in real time, such as seeing things in slow motion (e.g., the motion of a wave) or speeded up (e.g., erosion caused by a river). It is also useful when it is necessary to freeze action or replay it.
- For modeling scientific phenomena that are invisible to the naked eye (e.g., the movement of molecules in a gas).
- For working safely in lab-like simulations that would otherwise be hazardous (e.g., using dangerous chemicals) or messy in an assessment situation.
- For situations that require several repetitions of an experiment in limited assessment time, while varying the parameters (e.g., rolling a ball down a slope while varying the mass, the angle of inclination, or the coefficient of friction of the surface).
- For searching the Internet and resource documents that provide high-fidelity situations related to the actual world in which such performances are likely to be observed.
- For manipulating objects in a facile manner, such as moving concept terms in a concept map.

Quellmalz et al. describe a variety of other innovative assessments using ICT to increase both the sophistication of the problems posed to students and the observations the measures collect and analyze to make inferences about what students know and can do (Quellmalz, Timms, & Schneider 2009). An initiative likely to add to these innovative strategies is the Cisco-Intel-Microsoft project on the Assessment and Teaching of 21st Century Skills (http://www.atc21s.org). This effort plans both to define 21st-century skills in ways amenable to measurement and to develop technology-based assessments that are reliable and valid for these sophisticated intellectual and psychosocial skills.

Virtual Assessments: What Leading-Edge Technology Can Now Offer

The troubling findings about performance assessments are largely due to the intrinsic constraints of paper-based measures, coupled with the limited capabilities of virtual assessments, based on what computers and telecommunications could accomplish a decade ago, when research on performance assessments in accountability settings flourished. Assessments developed with current, more sophisticated immersive technologies face fewer threats from generalizability and efficiency concerns than traditional performance assessments. By alleviating dependence issues that arise from missing some subpart of the task (i.e., via simply putting the student on the right track after a wrong answer or a certain time limit), comparable generalizability can be demonstrated with significantly fewer tasks than traditional performance assessments. In addition, because virtual situations are more easily uniformly replicated, virtual performance assessments may not encounter the significant error from occasion. As mentioned above, studies on traditional performance assessments found that student performances varied on one testing occasion to another (Cronbach et al., 1997) and that the occasion of the sample was compounded with task sampling (Shavelson et al., 1999). Virtual assessments can be more cost effective as well as easier to administer and score for schools, and it can address task and occasion sampling variability through design. For example, as opposed to kits of tasks that contain items and objects, virtual performance assessments enable automated and standardized delivery via a Web-based application. By making the subtasks independent, one can design assessments with a larger number of tasks, thus increasing the reliability of the instrument.

Feasibility of Virtual Environments

Sophisticated educational media, such as single-user and multi-user virtual environments, extend the nature of the performance challenges presented and the knowledge and cognitive processes assessed. Single-user and multi-user virtual environments (MUVEs) provide students with an "Alice in Wonderland" experience: Students have a virtual representation of themselves in the world, called an "avatar," that one may think of as a digital puppet. This avatar enters the "looking glass" (monitor screen) to access a 3-D virtual world. These simulated contexts provide rich environments in which participants interact with digital objects and tools, such as historical photographs or virtual microscopes. Moreover, this interface facilitates novel forms of communication between students and computer-based agents using media such as text chat and virtual gestures (Clarke, Dede, & Dieterle, 2008). This type of "mediated immersion" (pervasive experiences within a digitally enhanced context) enables instructional designers to create curricula that are intermediate in complexity between the real world and simple structured exercises in K–12 classrooms. These new technologies allow instructional designers to construct both individual and shared simulated experiences that are otherwise impossible in school settings.

For almost a decade, we have been studying the feasibility and practicality of using MUVEs to increase student achievement in scientific inquiry (Dede, 2009). In this research, we have studied how virtual environments enable students to do authentic inquiry and engage in the processes of science. We have conducted a series of quasi-experimental design studies to determine if virtual environments can simulate real-world experimentation and provide students with engaging, meaningful learning experiences that increase achievement in scientific inquiry. Our results from a series of research studies show that these virtual environments enable students to engage in authentic inquiry tasks (problem finding and experimental design) and increase students' engagement and self-efficacy (Clarke & Dede, 2007; Clarke, Dede, Ketelhut, & Nelson, 2006; Ketelhut, 2007; Nelson, Ketelhut, Clarke, Bowman, & Dede, 2005; Nelson, 2007).

We have also found that student performance on the multiple-choice posttests do not necessarily reflect learning that we see via interviews, observations, summative essays, and analyses of log file data that capture students' activity as they interact with the environment (Clarke, 2006; Ketelhut, Dede, Clarke, Nelson, & Bowman, 2008; Ketelhut & Dede, 2006). We have built rich case studies of student learning in which we triangulate and compare different data sources, both qualitative and quantitative, to illustrate and understand students' inquiry learning (Clarke, 2006; Clarke & Dede, 2005; Clarke & Dede, 2007; Ketelhut et al., 2008). As a finding of our studies on how virtual environments foster inquiry learning, we have established that the multiple-choice assessments we use, even after extensive refinement, do not fully capture students' learning of inquiry skills.

Overall, our and others' studies of these virtual environments (Barab, Thomas, Dodge Carteaux, & Tuzun, 2004; Neulight, Kafai, Kao, Foley, & Galas, 2007) show that this type of experiential medium is a potential platform for providing both single and multi-user virtual performance assessments. Further, the fact that this medium underlies rapidly growing entertainment applications and environments such as multi-player Internet games (e.g., World of Warcraft, America's Army) and "virtual places" (e.g., Second Life) ensures the continuing evolution of sophisticated capabilities and authoring systems for these virtual environments.

In our River City research, we studied how MUVEs can provide students with authentic, situated learning experiences characteristic of the inquiry-based instruction proposed by the NRC, National Science Teachers Association (NSTA), and the American Association for the Advancement of Science (AAAS) standards. We found that MUVE environments and similar interactive, immersive media enable the collection of very rich observations about individual learners that provide insights on how students learn and engage in the inquiry processes (Clarke, 2009b; Ketelhut et al., 2008). Research on game-like simulations for fostering student learning is starting to proliferate, yet studying the potential of this medium for assessing student learning in a standardized setting is as yet untapped. We believe that virtual performance assessments developed with immersive technologies have the potential to provide better evidence for assessing inquiry processes. A decade ago, the implications from Shavelson and his colleagues' work were that the field needed multiple formats and multiple measures. This can be done seamlessly with immersive media, which allow for combining multiple modes and triangulating the results to provide an argument for student learning.

Using Immersive Technologies for Performance Assessment

With funding from the Institute of Education Sciences (IES), we are developing and studying the feasibility of virtual performance assessments to assess scientific inquiry for use in school settings as a standardized component of an accountability program (see Clarke, 2009a; http://www.virtualassessment.org). The research questions we are addressing in this project are:

- Can we construct a virtual assessment that measures scientific inquiry, as defined by the National Science Education Standards (NSES)? What is the evidence that our assessments are designed to test NSES inquiry abilities?
- Are these assessments reliable?

Design Framework

As stated earlier, the goal of an assessment is to provide valid inferences related to particular expectations for students (Linn et al., 2002). During the past decade, researchers have made significant advances in methods of assessment design. Frameworks such as the Assessment Triangle (NRC, 2001) and Evidence-Centered Design (ECD) (Mislevy Steinberg, & Almond, 2003; Mislevy & Haertel, 2006) provide rigorous procedures for linking theories of learning and knowing to demonstrations to interpretation. As mentioned above, we are using the ECD framework (Mislevy, Steinberg, & Almond, 2003; Mislevy & Haertel, 2006) to ensure construct validity of the assessments we develop.

ECD is a comprehensive framework that contains four stages of design:

1. Phase 1: domain analysis
2. Phase 2: domain modeling
3. Phase 3: conceptual assessment framework and compilation
4. Phase 4: a four-phase delivery architecture

Phases 1 and 2 focus on the purposes of the assessment, nature of knowing, and structures for observing and organizing knowledge.

Phase 3 is related to the Assessment Triangle. In this stage, assessment designers focus on the student model (what skills are being assessed), the evidence model (what behaviors/performances elicit the knowledge and skills being assessed), and the task model (what situations elicit the behaviors/evidence). Like the Assessment Triangle, these aspects of the design are interrelated. In the compilation stage of Phase 3, tasks are created. The purpose is to develop models for schema-based task authoring and developing protocols for fitting and estimation of psychometric models.

Phase 4, the delivery architecture, focuses on the presentation and scoring of the task (Mislevy et. al. 2003, Mislevy & Haertel, 2006).

According to White and colleagues, scientific inquiry is an active process comprised of four primary components: (a) theorizing, (b) questioning and hypothesizing, (c) investigating, and (d) analyzing and synthesizing (White & Frederiksen, 1998; White, Frederiksen, & Collins, in preparation). Measuring these inquiry processes as well as the products that result from the processes have long been challenges for educators and researchers (Marx et al., 2004). We have reframed White and Frederiksen's four categories into knowledge, skills, and abilities that we want to assess. We linked these skills and abilities back to the NSES and the NAEP framework. Starting with a large list of skills, we narrowed them down via working through the ECD

Figure 2. Authentic Alaskan bay.

framework and developing tasks that would allow us to elicit evidence that students are identifying a problem, using data to provide evidence, and interpreting evidence.

Virtual Assessments

The assessments we are creating are simulations of real ecosystems with underlying causal models. With immersive simulations, we can vary our causal models or can alter the variables to create extreme conditions under which students can conduct an experiment. Our first assessment is based on a real bay in Alaska. We took a real ecosystem (see Figure 2) and created a high-fidelity immersive virtual environment (see Figure 3).

In this assessment, students investigate the marine ecosystem and must discover why the kelp forest is depleting. They take on the identity of a scientist and have an avatar they move around the virtual world (see Figure 4, p. 322). Their avatar can interact with nonplayer characters (NPCs) we have programmed; walk around and observe visual clues; and use tools of scientists to measure salinity, nitrates, and temperature anywhere in the world (see Figure 5, p. 323).

Before students enter the environment, they watch a video that introduces them to the narrative. The first part of the assessment borrows from game design and sends students on quests that lead them to make observations and inferences about the kelp. They spend time gathering information and then are asked to identify the problem. These early stages of inquiry, such as data gathering and problem identification, are difficult to capture in a multiple-choice test but easily captured via a student's movement and actions in the world. As part of the interactivity of walking around the world, students can interact with different residents who inhabit the bay. For

Figure 3. High-fidelity virtual assessment.

example, there is a Park Ranger, an NPC that we have programmed, who is the initial quest giver and contact in the bay for the student. The other NPCs are located throughout the bay and provide hints or distracters for what is causing the problem. For example, at the wharf are fishermen, a hiker, and a power plant employee. These NPCs will respond to multiple questions, but students must select only one or two questions to ask. Thus, students have to make decisions about what questions to ask people in addition to what type of tests they need to conduct to gather data. These actions are all recorded and provide observations about students' inquiry processes.

However, we are not simply capturing or counting clicks. We are creating ways to triangulate performances by having students provide feedback on why they collected data or why they made a particular choice. We are developing performance palettes as an interactive method for assessing student product and performance (see Figures 6 and 7, pp. 324–325). We are utilizing current technologies that will allow us to rely less on text and more on interactive representations such as "drag-and-drop" text interfaces and audio queries.

We are creating assessments that allow students to spend time investigating a problem space and engage in the inquiry process. Unlike previous performance assessments vulnerable to a student making a mistake in the early stages of a task, we can make sure students always have the information they need to perform the task at hand. In a student's inquiry process, s/he moves from problem identification, to hypothesizing and questions, and ultimately to investigating. Through narrative, we can continually update students with accurate information they need to progress through the different phases of inquiry, such as problem identification and hypothesizing, thereby ensuring various phases of assessing their understanding are independent, not interdependent.

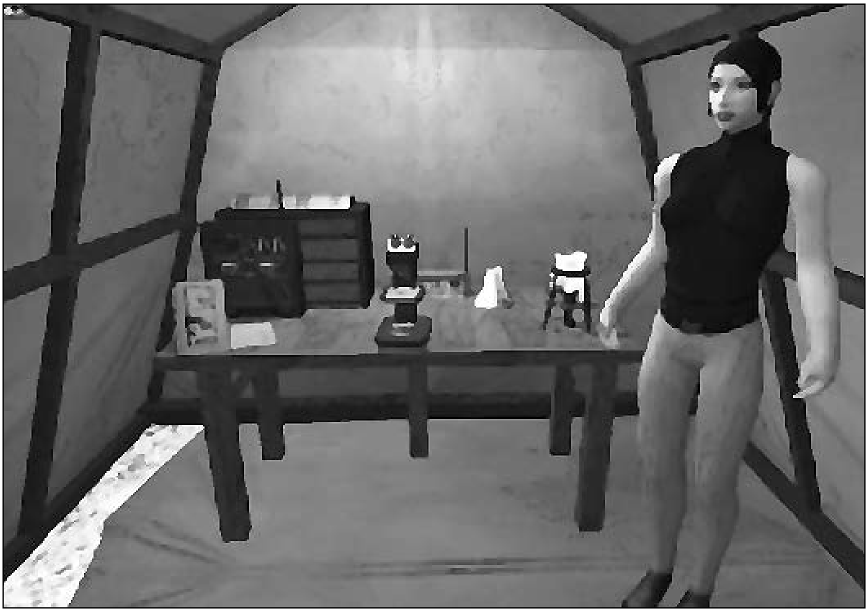


Figure 4. Avatar.

In the real world, allowing students to conduct an experiment has limitations. In a virtual setting, we can add as many constraints we want or leave the problem space open. Students are able to choose what kind of change they want to make within the bay (shut down a power plant that is dumping a substance into the bay, build barriers to prevent runoff from going into the water, plant trees, etc.) and then make that change. They then go into the environment and investigate how that change effected the kelp.

We are capturing everything the student does in the world, from the moment they enter the environment until the moment they leave. These data streams are recorded in XML in a back-end architecture that allows for real-time analysis of student paths in the ecosystem as well as logging for later analysis. Through our use of XML, the data are tokenized; that is, sensitive data such as student responses or test scores are shielded or kept out of the data stream to substantially minimize any risk of compromise or exposure. To further ensure data integrity, we encrypt them before sending.

Method

Validity is a central issue in test construction. According to Messick (1989), "Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment…" (cited in Messick, 1994, p. 13). To provide evidence that our assessment questions test students' ability to do inquiry as outlined by the NSES, we plan to conduct a series of validity studies that result in evidence

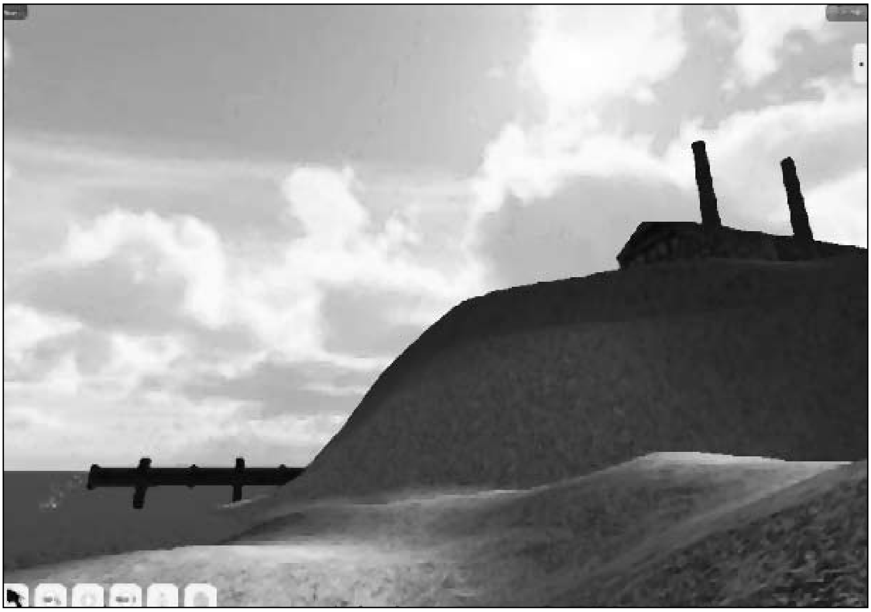


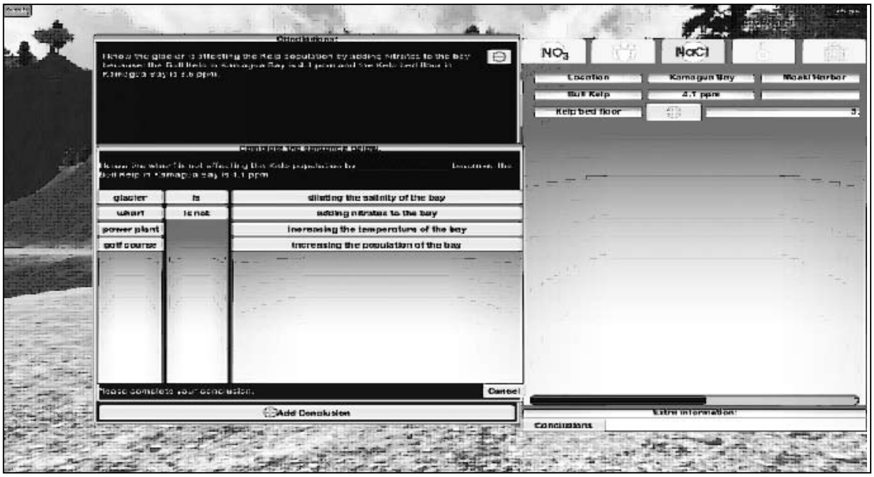
Figure 5: Visual clues.

on construct validity. We will employ similar methods to those carried out in the Validities of Science Inquiry Assessments (VSIA) (Quellmalz, Kreikemier, DeBarger, & Haertel, 2006). We will conduct both an alignment analysis and a cognitive analysis of our assessments, because these methods provide valuable, separate sources of validity evidence (Quellmalz et al., 2006).

To assess the reliability of performance assessments, we plan to conduct a series of generalizability studies. Generalizability theory (g-theory) is a statistical theory that allows decision makers to study the dependability of behavioral measures and procedures (Shavelson & Webb, 1991). It is a commonly used technique for making decisions and drawing conclusions about the dependability of performance assessments (Baxter, 1995; Baxter & Shavelson, 1994; Pine et al., 1993; Rosenquist, Shavelson, & Ruiz-Primo, 2000; Shavelson, Baxter, & Pine, 1991). G-theory was first introduced as a statistical theory by Cronbach, Gleser, Nanda, and Rajaratnam (1972) to extend the limitations of using classical test theory, which provides an estimate of a person's true score on a test, by allowing researchers to generalize about a persons' behavior in a defined universe from observed scores (Shavelson, Webb, & Rowley, 1989). Further, classical test theory can estimate only one source of error at a time (Cronbach et al., 1972; Shavelson & Webb, 1991), whereas in g-theory, multiple sources of error can be measured in a single analysis.

Conclusion

The assessments we are creating will complement rather than replace existing standardized measures by assessing skills not possible via item-based



Conclusions:

I know the glacier is affecting the Kelp population by adding nitrates to the bay because the Bull Kelp in Kamagua Bay is 4.1 ppm and the Kelp bed floor in Kamagua Bay is 5.6 ppm.

Complete the sentence below:

I know the wharf is not affecting the Kelp population by ________ because the Bull Kelp in Kamagua Bay is 4.1 ppm

glacier	is	diluting the salinity of the bay
wharf	is not	adding nitrates to the bay
power plant		increasing the temperature of the bay
golf course		increasing the population of the bay

Please complete your conclusion.

Location	Kamagua Bay	Moaki Harbor
Bull Kelp	4.1 ppm	
Kelp bed floor		5.

Extra information:

Conclusions

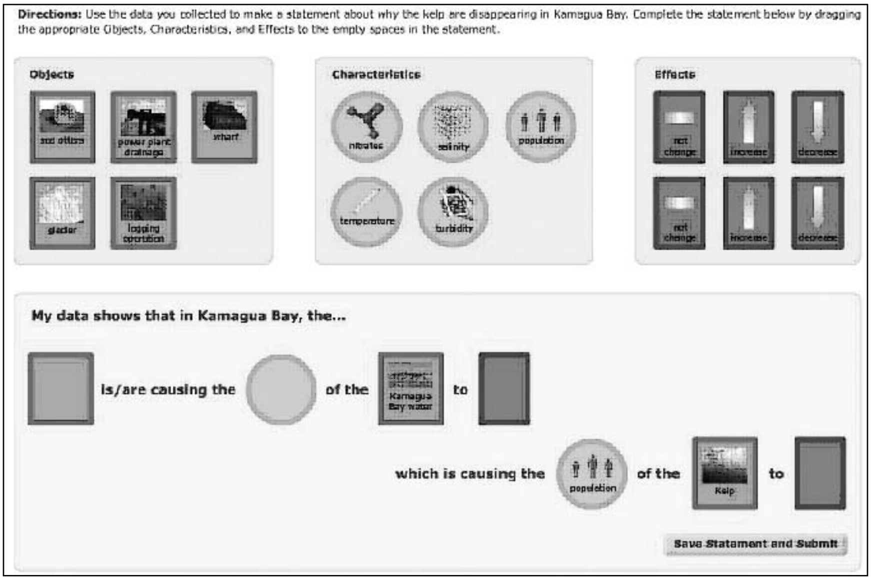
Figure 6. Embedded performance palette.

paper-and-pencil tests or hands-on real-world performance assessments. One of the advantages of developing virtual assessments is that they will alleviate the need for extensive training for administering tasks. It is difficult to standardize the administration of paper-based performance assessments, and extensive training is required to administer the tasks. With virtual assessments, we can ensure standardization by delivering instruction automatically via the technology.

A second advantage is that virtual assessments alleviate the need for providing materials and kits for hands-on tasks. Everything will be inside the virtual environment. Third, these performance assessments will be easier to administer and will require very little, if any, training of teachers. Scoring will all be done behind the scenes; there will be no need for raters or training of raters. Fourth, virtual assessments would alleviate safety issues and inequity due to lack of resources

Further, using digital media for assessment enables the use of measures based on performance using visualizations, simulations, data-analysis tools, and GIS representations. Virtual assessments will allow us to include visualization of data and information, including phenomena that can't be observed with the naked eye or even in real time.

In our work in developing virtual inquiry curricula, we developed the ability to simulate the passing of time, to allow students to collect data on change over time, and to conduct experiments where time can be fast-forwarded. These capabilities allow for rich learning experiences and the ability to conduct experiments that may take too much time to use for learning or assessment purposes. The potential to develop assessments that can condense time and experiments within a class period opens new doors for assessing inquiry and students' ability to conduct empirical investigations.



Directions: Use the data you collected to make a statement about why the kelp are disappearing in Kamagua Bay. Complete the statement below by dragging the appropriate Objects, Characteristics, and Effects to the empty spaces in the statement.

My data shows that in Kamagua Bay, the...

is/are causing the **of the** Kamagua Bay water **to**

which is causing the population **of the** Kelp **to**

Figure 7. External performance palettes.

For example, virtual assessments can allow for interactive speeds (slow, fast, rewind) and the ability to show change over time quickly, repeat steps, or vary parameters. The advances in technologies for teaching concepts like Newtonian physics could also be used to assess students' understanding of such concepts.

Whereas our current work centers on the potential of immersive media to capture rich observations of student learning, other types of ICT allow for similarly rich observations. For example, wikis and other forms of Web 2.0 media, asynchronous discussions, intelligent tutoring systems, games, and augmented realities all provide potential for capturing mediated experiences. In the future, we hope the field can build on our work to explore these other possibilities for collecting and analyzing rich data about student engagement and learning.

Author Notes

Jody Clarke-Midura is a research associate at Harvard University. Her research focuses on how immersive technologies enable new approaches and methods for assessing and understanding learning. Correspondence concerning this article should be addressed to Jody Clarke-Midura, Larsen Hall, 14 Appian Way, Cambridge, MA 02138. E-mail: Jody_Clarke@mail.harvard.edu

Chris Dede is the Timothy E. Wirth Professor in Learning Technologies at Harvard's Graduate School of Education. His fields of scholarship include emerging technologies, policy, and leadership. Correspondence concerning this article should be addressed to Chris Dede, Longfellow Hall, 13 Appian Way, Cambridge, MA 02138. E-mail: Chris_Dede@harvard.edu

References

Aspen Institute. (2007). *Beyond NCLB: Fulfilling the promise to our nation's children.* Washington, DC: Aspen Institute.

Barab, S., Thomas, M., Dodge, T., Carteaux, R., & Tuzun, H. (2004). Making learning fun: Quest Atlantis, a game without guns. *Educational Technology Research & Development, 53*(1), 86–108.

Baxter, G. P. (1995). Using computer simulations to assess hands-on science learning. *Journal of Science Education and Technology, 4*, 21–27.

Baxter, G. P., & Shavelson, R. J. (1994). Science performance assessments: Benchmarks and surrogates. *International Journal of Educational Research, 21*, 279–298.

Clarke, J. (2006). *Making learning meaningful: An exploratory study of multi-user virtual environments in middle school science.* Qualifying Paper submitted to the Harvard Graduate School of Education. Cambridge, MA.

Clarke, J. (2009a). *Studying the potential of virtual performance assessments for measuring student achievement in science.* Paper presented at the Annual Meeting of the American Educational Research Association (AERA), San Diego, CA. Retrieved December 30, 2009, from http://virtualassessment.org/publications/aera_2009_clarke.pdf

Clarke, J. (2009b). *Exploring the complexity of learning in an open-ended problem space.* Doctoral dissertation submitted to Harvard Graduate School of Education, Cambridge, MA.

Clarke, J., & Dede, C. (2005). *Making learning meaningful: An exploratory study of using multi-user environments (MUVEs) in middle school science.* Paper presented at the Annual Meeting of the American Educational Research Association, Montreal, Canada. Retrieved December 30, 2009, from http://muve.gse.harvard.edu/rivercityproject/documents/aera_2005_clarke_dede.pdf

Clarke, J. & Dede, C. (2007). MUVEs as a powerful way to study situated learning. In C. A. Chinn, G. Erkens, & S. Putambekar (Eds.), *The Proceedings of Conference for Computer Supported Collaborative Learning (CSCL)* (pp. 141–144). Mahwah, NJ: Erlbaum.

Clarke, J., Dede, C., & Dieterle, E. (2008). Emerging technologies for collaborative, mediated, immersive learning. In J. Voogt & G. Knezek (Eds.), *The international handbook of technology in primary and secondary education* (pp. 901–910). New York: Springer-Verlag.

Clarke, J., Dede, C., Ketelhut, D. J., & Nelson, B. (2006) A design-based research strategy to promote scalability for educational innovations. *Educational Technology, 46*(3), 27–36.

Confrey, J., & Maker, K. M. (2005). Critiquing and improving the use of data from high-stakes tests with the aid of dynamic statistics software. In C. Dede, J. P. Honan, & L. C. Peters (Eds.), *Scaling up success: Lessons from technology-based educational improvement* (pp. 198–226). New York: Wiley.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: Wiley.

Cronbach, L. J., Linn, R. L., Brennan, R. L, & Haertel, E. H. (1997). Generalizability analysis for performance assessments of student achievement or school effectiveness. *Educational and Psychological Measurement, 57*, 373–399.

Dede, C. (2007). Reinventing the role of information and communications technologies in education. In L. Smolin, K. Lawless, & N. Burbules (Eds.), *Information and communication technologies: Considerations of current practice for teachers and teacher educators.* NSSE Yearbook 2007, 106(2), 11–38. Malden, MA: Blackwell Publishing.

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

Dede, C. (in press). Technological supports for acquiring 21st-century skills. In E. Baker, B. McGaw, & P. Peterson (Eds.), *International encyclopedia of education* (3rd ed.). Oxford, England: Elsevier.

Haertel, G. D., Lash, A., Javitz, H., & Quellmalz, E. (2006). *An instructional sensitivity study of science inquiry items from three large-scale science examinations.* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Ketelhut, D. J. (2007). The impact of student self-efficacy on scientific inquiry skills: An exploratory investigation in River City, a multi-user virtual environment. *The Journal of Science Education and Technology, 16*(1), 99–111.

Ketelhut, D. J., & Dede, C. (2006). *Assessing inquiry learning.* Paper presented at the National Association of Research in Science Teaching, San Francisco, CA.

Ketelhut, D., Dede, C., Clarke, J., Nelson, B., & Bowman, C. (2008). Studying situated learning in a multi-user virtual environment. In E. Baker, J. Dickieson, W. Wulfeck, & H. O'Neil (Eds.), *Assessment of problem solving using simulations* (pp. 37–58). Mahweh, NJ: Erlbaum.

Linn, R. L. (1994). Performance assessment: Policy promises and technical measurement standards. *Educational Researcher, 23*(9), 4–14.

Linn, R. L. (2000). *Assessments and accountability. Educational Researcher, 29*(2), 4–16.

Linn, R. L., Baker, E. L., & Betebenner, D. W. (2002). Accountability systems: Implications of requirements of the "'No Child Left Behind Act of 2001." *Educational Researcher, 31*(6), 3–26.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*(8), 15–21.

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

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: Macmillan.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13–23.

Mislevy, R., & Haertel, G. (2006). *Implications of evidence-centered design for educational testing* (Draft PADI Technical Report 17). Menlo Park, CA: SRI International.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1,* 3–62.

National Assessment Governing Board. (2004). NAEP 2009 science framework development: Issues and recommendations. Washington, D.C.: Author.

National Assessment Governing Board. (2007). *Science framework for the 2009 National Assessment of Educational Progress.* Washington, D.C.: Author.

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment.* Washington, D.C.: National Academy Press.

National Research Council. (2006). S*ystems for state science assessment.* Washington, D.C.: The National Academies Press.

Nelson, B. (2007). Exploring the use of individualized, reflective guidance in an educational multi-user virtual environment. *The Journal of Science Education and Technology, 16*(1) 83–97.

Nelson, B., Ketelhut, D. J., Clarke, J., Bowman, C., & Dede, C. (2005). Design-based research strategies for developing a scientific inquiry curriculum in a multi-user virtual environment. *Educational Technology, 45*(1), 21–27.

Neulight, N., Kafai, Y. B., Kao, L., Foley, B., & Galas, C. (2007). Children's learning about infectious disease through participation in a virtual epidemic. *Journal of Science Education and Technology 16*(1), 47–58.

Nicols, S. L., Glass, G. V., & Berliner, D. C. (2005). High stakes testing and student achievement: Problems for the No Child Left Behind Act. Tempe, AZ: Educational Policy Research Unit, Arizona State University.

Pine, J., Baxter, G., & Shavelson, R. (1993). Assessments for hands-on elementary science curricula. MSTA Journal, 39(2), 3, 5–19.

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. Washington, DC: National Research Council.

Quellmalz, E., Kreikemeier, P., DeBarger, A. H., & Haertel, G. (2006). *A study of the alignment of the NAEP, TIMSS, and New Standards Science Assessments with the inquiry abilities in the National Science Education Standards.* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

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

Quellmalz, E. S., Timms, M. J., & Schneider, S. A. (2009). *Assessment of student learning in science simulations and games.* Washington, D.C.: National Research Council.

Resnick, L. B., & Resnick, D. P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B. Gifford & M. O'Connor (Eds.), *Changing assessments: Alternative views of aptitude, achievement, and instruction* (pp. 37–75). Norwell, MA: Kluwer Academic Publishers.

Rosenquist, A., Shavelson, R. J., & Ruiz-Primo, M. A. (2000). *On the "exchangeability" of hands-on and computer simulation science performance assessments.* Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing, Center for the Study of Evaluation, UCLA.

Salomon, G. (Ed.). (1993) *Distributed cognitions: Psychological and educational considerations.* New York: Cambridge University Press.

Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement, 30*(3), 215–232.

Shavelson, R. J., Baxter, G. P., & Pine, J. (1991). Performance assessment in science. *Applied Measurement in Education, 4*(4), 347–362.

Shavelson, R. J., Ruiz-Primo, M. A., & Wiley, E. W. (1999). Note on sources of sampling variability in science performance assessments. *Journal of Educational Measurement, 36*, 61–71.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer.* Thousand Oaks, CA: Sage.

Shavelson, R. J., Webb, N. M., & Rowley, G. (1989). Generalizability theory. *American Psychologist, 44*, 922–932.

Stecher, B. M., & Klein, S. P. (1997). The cost of science performance assessments in large-scale testing programs. *Educational Evaluation and Policy Analysis, 19*(1), 1–14.

White, B., Collins, A., & Frederiksen, J. (under review). The nature of scientific meta-knowledge. *International Journal of the Learning Sciences.*

White, B., & Frederiksen, J. (1998). Inquiry, modeling, and metacognition: Making science accessible to all students. *Cognition and Instruction, 16*(1), 3–118.