CRESST REPORT 735

Terry P. Vendlinski
Eva L. Baker
David Niemi

TEMPLATES AND OBJECTS IN AUTHORING PROBLEM-SOLVING ASSESSMENTS

MAY, 2008

National Center for Research on Evaluation, Standards, and Student Testing

Graduate School of Education & Information Studies
UCLA | University of California, Los Angeles

Templates and Objects in Authoring Problem-Solving Assessments

CRESST Report 735

Terry P. Vendlinski, Eva L. Baker, & David Niemi
CRESST/University of California, Los Angeles

May, 2008

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Center for the Study of Evaluation (CSE)
Graduate School of Education & Information Studies
University of California, Los Angeles
300 Charles E. Young Drive North
GSE&IS Bldg., Box 951522
Los Angeles, CA 90095-1522
(310) 206-1532

The work reported herein was supported under the Inter-agency Educational Research Initiative, Award Number 0129406.

TEMPLATES AND OBJECTS IN AUTHORING PROBLEM-SOLVING ASSESSMENTS[1,2]

Terry P. Vendlinski, Eva L. Baker, and David Niemi
CRESST/University of California, Los Angeles

Abstract

Assessing whether students can both re-present a corpus of learned knowledge and also demonstrate that they can apply that knowledge to solve problems is key to assessing student understanding. This notion, in turn, impacts our thinking about what we assess, how we author such assessments, and how we interpret assessment results. The diffusion of technology into venues of learning offers new opportunities in the area of student assessment. Specifically, computer-based simulations seem to provide sufficiently rich environments and the tools necessary to allow us to infer accurately how well a student's individual mental model of the world can accommodate, integrate, and be used to exploit concepts from a domain of interest. In this paper then, we first identify the characteristics of simulations that our experience suggests are necessary to make them appropriate for pedagogical and assessment purposes. Next, we discuss the models and frameworks (templates) we have used to ensure these characteristics are considered. Finally, we describe two computerized instantiations (objects) of these frameworks and implications for the follow-on design of simulations.

Introduction

The idea that students should understand the concepts they are learning well enough to actually apply them in an appropriate context has been a central theme in both teaching and educational assessment for more than a century (Gould, 1996). Although achieving this ideal has remained elusive for most of that century, the growing presence of computers in U.S. classrooms seems to offer us an opportunity now to make applied problem solving the norm in many educational domains and to change how such exercises are evaluated and used by educators and policymakers (Baker, 2004; Edelstein, Reid, Usatine, & Wilkes, 2000). In fact, educational stakeholders are increasingly asking that students not only demonstrate that they can re-present a corpus of learned knowledge but also demonstrate the reasoning necessary to apply that knowledge to solve problems likely to be faced in future educational or other life pursuits (Herman, 1992; Quellmalz & Hartel, 2004). To exploit these new opportunities fully

[1]This report was first printed as chapter 16 of *Assessments of Problem Solving Using Simulations* (Vendlinski, Baker, & Niemi, 2008).

[2]We would like to thank Dr. William Bewley and Dr. Greg Chung for their feedback during the preparation of this chapter and Joanne Michiuye and Bryan Hemberg for editing our work.

and satisfy the demands of these stakeholders requires new ways of thinking about what we assess, how we author such assessments, and how we interpret assessment results.

What has not seemed to change in this discussion is what we mean by the terms problem and problem solving. In this report, we take as our definition of a problem the one proffered by Newell and Simon (1972): A person recognizes they have a problem when they desire a goal but do not immediately know the series of actions necessary to achieve the goal. Problem solving, then, is the series of mental or physical actions a solver takes to transform the present state (desiring to achieve the goal) to the final state (achieving the goal). For now, we make the assumption that the problem solver's goal is identical to the one intended by the problem (assessment) designer, but we relax this assumption when inferring goals from the actual solution strategies used by a problem solver.

Although the physical actions of a problem solver are observable, in most educational settings we must usually infer the mental activity either from the physical activity itself, from the problem solver self-reports, or from some combination of the two. Previously, evaluations of student problem solving relied almost exclusively on self-reports provided by the student. Evaluating such written or oral reports of activity not only imposed a time burden on evaluators but also often introduced variables such as student writing ability and self-filtering into our inferences (Mayer, 2003). An additional difficulty in inferring mental action from physical activity was our inability to provide a rich enough problem-solving space to accommodate the support (scaffolding) or tools that a problem solver needed. Because such limitations can have dramatic effects on how students solve problems, they can lead to faulty inferences about student understanding (Gobert, Buckley, & Clarke, 2004; Norman, 1993; Rogoff, 1998). Our challenge, then, is to provide sufficiently rich environments and necessary tools that will allow us to infer accurately how well a student's individual mental model of the world can accommodate, integrate, and be used to explain concepts from the domain of interest (Buckley & Boulter, 2000; Seel & Schenk, 2003).

Modern computer-based simulations offer an opportunity to meet this challenge. First, they allow us to offer test takers a large, but finite, problem-solving space with a given amount of complexity. Second, they allow us to record every interaction a student has with tools in the problem space and how (or if) the student uses each tool. When used in a controlled environment (such as a classroom or under observation), we can further control or account for the external tools and artifacts to which problem solvers have access when they attempt to solve the problem. The actions a student takes to solve a particular problem should allow us to make valid inferences about how a student couples available tools with extant understanding and the depth of that understanding (Vendlinski, 2001; Vendlinski & Stevens,

2000). Consequently, assessment designers can both focus on specific solution strategies, if desired, and limit the complexity of the problem space.

It is these limitations in problem space complexity that give rise to simulations of reality. As an imitation of reality, simulations are designed to replicate the real world or have the appearance of reality without the same complexity, cost, danger, or inaccessibility. Simulations are to reality what the small-scale map is to a 1:1 scale map of the world. As in the real world, however, we would also require that simulations have some capacity for problem solvers to formulate and test various hypotheses or to follow various paths to reach a conclusion and that there be feedback to the problem solver as the solver makes perturbations in the system. Because well-tailored feedback has been consistently shown to improve educational outcomes (Black & Wiliam, 2004), simulations might logically promote the same outcomes.

We view simulations as a subset of problem-solving assessment environments; others include, but are not limited to, written or verbal applications of knowledge to a given situation (Baker, Freeman, & Clayton, 1991); explanations of a problem and proposed solution (Schworm & Renkl, 2006); and symbolic or written explanations of problem solutions, including worked examples (see, e.g., Halabi, Tuovinen, & Farley, 2005; Paas, Renkl, & Sweller, 2003; Renkl, 2002; Sweller & Cooper, 1985). In each case, students are presented a particular problem context, and they apply their extant knowledge to reach a solution or desired goal. Collectively, we refer to such problem contexts as information sources, and we see simulations as a special category of information source because they allow students to interact with dynamic information that, through feedback, allows users to formulate or even reformulate problem solutions.

This conceptualization of simulations integrates well with the vision that learning is a transformation that allows someone capable of certain performances to become capable of performing additional ones (usually better) without losing preexisting ability and usually integrated with other capabilities so they can be evoked when appropriate (Newell & Simon, 1972). This presupposes, however, that simulations be used properly, namely, that they "help people of all ages make connections among different aspects of their knowledge" (Bransford, Brown, & Cocking, 1999, p. 92) by building on what they know and how we scaffold the task (Vygotsky, 1962). Among other activities, Bransford and colleagues (1999) suggested that this scaffolding includes interesting and motivating the child, adapting the task to the cognitive ability of the learner, and providing feedback. Although motivation is an important piece of this mix (Quinn, 2005), cognitive ability seems to be as important, and knowledge (organized domain-specific knowledge, self-regulation, and problem-solving strategies), in at

least one study, has been found to explain almost four times as much of the variance in student learning as motivation (Schraw, Brooks, & Crippen, 2005). The multimedia tools made possible by computer technology now make it possible to design intriguing situations that involve the learner in an inquiry process in which facts are gathered from data sources, similarities and differences among facts noted, and concepts developed.

Although simulations need not necessarily be digital and provide immediate feedback (e.g., actual wind tunnels in aerodynamics or multiyear plant propagation studies in horticultural genetics), one of the benefits of computerized simulations is that tools delivered on a computer can offer students immediate feedback as well as an opportunity for assessors to collect, organize, and analyze the voluminous amounts of data that often result from simulations. The computerized analytical tools this technology provides can also minimize evaluation time. Clearly, however, all computer-based information sources or problem spaces are not created equal.

Our experience suggests that educators search for and select learning and assessment experiences for a wide variety of reasons. For example, in a study (Vendlinski, Niemi, & Wang, 2005), we asked teachers to share the problem spaces they use to teach and assess the concepts of force and motion. Among the various ideas shared was a simulation task that required students to design a roller coaster from the cardboard tubes inside bathroom tissue. After a brief explanation, the teacher confided that she only required students to design a roller coaster that physically constrained a marble through a minimum number of turns. Although the activity was reportedly motivational and engaging for students, the teacher could identify no specific force and motion concepts, standards, or ideas that students needed to master or explain to complete this task successfully.

The overriding requirement that learning experiences be fun and motivational, regardless of their ability to stimulate or assess understanding of core concepts or "big ideas," is a consistent theme in the pedagogical situations we encounter. Although we concur that the motivation (desire) to reach a goal is necessary for problem solving and learning to occur (Caine R. N., Caine, G., McClintic, & Klimek, 2005; Dweck, 2002; Zull, 2002), it alone is not sufficient and ultimately, as suggested, may not be as important as how concepts are organized or how a problem solver reflects on his or her solution strategies. This suggests that it was not the information source (a simulation in the case of the marble roller coaster) that was faulty, but a mismatch between educational goals and a deficiency in the inferences made about student understanding based on the data provided by the simulation. The fact that students could design a track that constrained a rolling marble may have had little to do with the student's understanding of force and motion. As is the case with any type of assessment,

data do not equal valid inference. The purpose of the assessment and the inferences that will be drawn must be considered when selecting or designing assessments, including problem-solving simulations.

We have also been involved with educators who have developed learning experiences that both stimulated learning and provided valuable data on which to base summative and formative inferences of how well students had learned and could actually apply concepts. In one instance, for example, civil engineering students were asked to carry out an investigation of a hazardous waste site at an abandoned airfield. The Integrated Site Investigation Software (ISIS) simulation allowed students to develop links between classroom theory and real-world situations and to apply and test these theories. The simulation was embedded within instruction and was comprehensive in terms of both subject matter and the broader context of engineering. Results from the ISIS study suggested not only that student content understanding improved, but also that the students learned concepts at a deeper level. In addition, the students felt ISIS was effective at improving their ability to handle complex projects, allowed them to link classroom theory with real-world applications, and improved their problem-solving performance (Chung, Harmon, & Baker, 2001). We have also investigated the effective use of simulations in middle school and high school science, mathematics, and postgraduate instruction on decision analysis.

These experiences suggest to us a way to assess learning effectively and a method to mediate design through the use of templates and manage inference validity by employing objects. In the remainder of this report, we first identify the characteristics of simulations that our experience suggests are necessary to make them appropriate for pedagogical and assessment purposes. Next, we discuss the models and frameworks (templates) we have used to ensure these characteristics are considered. Finally, we describe two computerized instantiations (objects) of these frameworks and implications for the follow-on design of simulations.

Important Characteristics of Good Simulations

There is growing evidence that students learn best when they are presented with academically challenging work that focuses on individual sense making and building the necessary strategies (skills) to solve problems within a domain (Chung & Baker, 2003; Fuchs et al., 2004). In particular, simulations have been shown to improve learning and to provide important insights into student learning if properly implemented (Gredler, 2004; Leemkuil, Jong, & Ootes, 2000; Randel, Morris, Wetzel, & Whitehill, 1992; Rieber, 2005). Unfortunately, as discussed here and in other chapters in the current volume, merely

engaging student interest is insufficient to motivate or assess deep conceptual understanding of a knowledge domain or to help students develop rich schema.

Good simulations (like other instructional materials and assessments) have proven difficult to integrate effectively with instruction for a number of reasons. Our experience suggests that effective use of problem-solving simulations and assessments require they:

Support a clearly stated learning goal that is aligned with the overall instructional goal and must be reasonably expected to produce results suitable to both developing and assessing the attainment of this goal. Validity of inference must be considered when designing and using any assessment, including simulation-based assessment (Gearhart et al., 2006). For example, we are aware of a simulation that was designed to increase and test student understanding of dissolved gases in the blood. The goal was clear and aligned with overall goals. The students, however, were able to complete the simulation by repeated guessing and checking their solution until they solved the problem. An e-mail to the students' instructor only required students to fill in the parameters they used to complete the simulation—parameters they could obtain directly from the simulation itself. Although a guess-and-check (generate-and-test) strategy might be an appropriate problem-solving methodology, one must be clear that the type of problem solving is clearly identified and can be validly inferred from simulation data. In this case, the instructor inferred that students who successfully completed the simulation had a deep understanding of blood gas chemistry, which was often an incorrect conclusion. Implicit in this requirement is that the content of the simulation or assessment is accurate.

Specify the degree of understanding necessary to solve the problem and allow assessors to accurately differentiate levels of understanding among problem solvers. We refer to this as the cognitive demand of the information source or assessment (Baker, 1998). Often, students with a wealth of prior knowledge or those who have been exposed to an identical (or nearly identical) problem before may be merely recalling a solution algorithm instead of developing a solution. This does not mean that simulations and assessments should never address recall, just that they must allow us to accurately discriminate deeper understanding from recall when such inferences are desired or required. A student may have to complete multiple (similar and dissimilar) cases of a simulation or assessment, and assessors might need to evaluate student solution strategies to make these types of inferences (Newell & Simon, 1972). As suggested by the expert-knowledge literature (see, e.g., Anderson & Leinhardt, 2002; Chi, Feltovich, & Glaser, 1981; Ericsson, 2003; Hmelo-Silver & Pfeffer, 2004; Jacobson, 2001), solution strategies should be indicative of the degree of student understanding represented. In fact, the Modeling Across the Curriculum project has reported (Buckley, Gobert, Gerlits, Goldberg, & Swiniarski, 2004) an ability to interpret student interactions with various simulations as evidence of specific student mental models and complex learning (or lack thereof). Implicit in this feature of simulation, however, is that it engage the student

sufficiently to ensure that we are measuring a student's cognitive ability and not the lack of motivation (Quinn, 2005).

Allow for alternative solution strategies indicative of understanding in the domain taught or assessed (Gredler, 2004). It is often assumed that students engaging in problem solving in a specific simulation or problem space will use a single, expected solution method. But, such a constrained process can hide alternative conceptions (including misconceptions) that more accurately represent a student's mental model, inform pedagogical interventions, and improve the accuracy of our inferences about student understanding (Glaser & Baxter, 2000; Stiggins, 1994). The key to identifying important deficiencies in student understanding, however, rests on our knowledge of the features of the performance that are most salient or indicative of the learning about which we want to make inferences (Bransford & Schwartz, 1999; Mayer, 2003). By targeting specific attributes of various solution strategies, we can infer a student's understanding of the concepts necessary to perform that task, including similarities to how experts organize the domain and the common misconceptions novices in that domain are likely to have (Cromley & Mislevy, 2004). If a simulation does not allow problem solvers to use various problem-solving strategies (both domain dependent and domain independent), then inferences about student understanding may be erroneous. Designers must consider these alternative strategies when developing simulation-based assessments and the evaluation methods they will use to draw inferences from the data these simulations generate. We have seen, for example, students correctly solve simulations because they were able to rule out all other answers based on the solutions to previous instantiations of the problem or students who used knowledge other than that provided in the simulation to narrow possible answers down to a meaningful few and then guess at an answer. Unfortunately, neither solution strategy was anticipated, and this led to the even more egregious inference that student understanding was improving as students accomplished each succeeding case of the simulation.

Have a level of complexity that corresponds to the learning goal. As problem spaces and assessments become more complex (either because the problem is virtually unbounded or because the number of items that the problem solver must consider is large), students may change the way they approach a problem (Kirschner, Sweller, & Clark, 2006). In particular, when the complexity of a problem space or a solution far exceeds student ability, students often change solution strategies or even give up on a solution (Vendlinski, 2001). This becomes a serious threat to the validity of our inferences about student understanding, especially when aspects of the simulation are irrelevant to the instructional goal (Mayer & Moreno, 2003). In such cases, an assessor may lose the ability to discern accurately what a student understands and may even discourage learning (Steffe & Thompson, 2000). We may also be unable to distinguish a lack of motivation to continue and an inability to continue that results from a lack of understanding. On the other hand, if the simulation is too easy, the simulation may not discern more sophisticated degrees of student understanding. Our experience, and that of

others (Mayer, 2003; Sweller, 2003), suggests that problem space complexity is partly related to the tools available to the problem solver and the number of items a student must remember. A related issue is that an overly constrained problem space may inaccurately represent the content domain and encourage the development of misunderstandings (VanLehn, 1990). Greer (1992), for example, noted that when children learn multiplication using only repeated integer addition, they often mistakenly conclude that the product of any multiplication must be larger than either the multiplier or the multiplicand. This raises consequential validity issues (Messick, 1989). An accurate task analysis should identify the constraints necessary to achieve the desired learning goal (Mislevy, Steinberg, Breyer, Almond, & Johnson, 1999; Steffe & Thompson, 2000).

Satisfy other practical constraints. The time necessary to administer and evaluate student performances on the simulation or assessment must be reasonable, and the differential impacts of time on subpopulations must be considered. Moreover, as solutions become more complex (involving many concepts) or if the steps to a solution can be ordered in many ways, the time required to make accurate evaluations of the ability of a test taker is further increased. As the complexity of the problem space increases, the resulting inferences about student understanding can become so nuanced and complex that they become unusable by teachers or policymakers. In addition to simulation attributes that detail how the simulation will interact with users, there are environmental considerations. Among these considerations are operating platform, software dependencies (including support applications such as Internet browsers, Adobe Acrobat or RealPlayer™), central processing unit speed, computer memory, fast interconnectivity, and technical support.

Using the Characteristics to Design Simulations and Assessments for Educational Use

A long history of research briefly recounted by DeCorte, Greer, and Vershaffel (1996) suggested educators have difficulty helping their students develop a deep understanding of concepts because prevalent instruction and assessment methods focus on the "recall of facts, computation, and standard procedures … [and] cannot yield useful information on problem solving, modeling of complex situations, or ability to communicate. … Nor can they provide the detailed diagnostic feedback for the teacher appropriate to the view of the learner as an individual constructor of knowledge" (p. 530). If our vision of education is for learners to problem solve, model, communicate, and develop higher-order thinking (National Council of Teachers of Mathematics [NCTM], 2000; Pellegrino, Chudowsky, & Glaser, 2001), then Glaser (2001) argued that, "Achievement measurement should be designed to emphasize not only content considerations, but also knowledge structures and process considerations that are involved in facilitating competence" (p. 19).

Along these lines, our current research (Vendlinski, Niemi, & Wang, 2005; Vendlinski, Niemi, Wang, & Monempour, 2005) suggests that we can scaffold instructional and assessment task design in a way that will encourage designers (a) to focus their assessment design on student understanding of the key concepts or principles that govern a domain (big ideas) rather than the recall of decontextualized facts; and (b) to develop scoring rubrics and methods as part of their assessment design and refinement that encourages reflection on the purpose of the assessment.

Various researchers have developed additional frameworks for assessment design using simulated tasks. One of the most researched of these models is Mislevy's idea of evidence-centered design (ECD). The research of Mislevy, Almond, and Lukas (2004) suggested an assessment framework that combines task and student models and so allows different classes of student responses to be aggregated with various statistical models to inform both instruction and learning theory. Mislevy and colleagues used ECD to examine real-world situations in which people engage in the behaviors and utilize the knowledge emblematic of a domain. They then determined the types of tasks appropriate for assessment, as well as performance features (including misconceptions) that may be important to capture in assessment. These tasks can then be modeled using templates. Furthermore, models of student cognition can be interpreted with probabilities of latent trait analysis and probabilistic (Bayesian) networks to determine student proficiency.

Tatsuoka K. K., Corter, and Tatsuoka C. (2004) focused on the rule-space method (RSM) to discover and measure important attributes of performance involved in domain competence. RSM develops "a one-to-one correspondence between subject item response patterns and the corresponding ideal item score patterns" (p. 905). For some years, brain and cognitive scientists have been investigating similar categorization schemes to explain how humans make meaning from a sensory input space (e.g., Bobick, 1987; Richards, Feldman, & Jepson, 1992).

Finally, Stevens and Palacio-Cayetano (2003) also developed a method for investigating student problem-solving strategies during scientific problem solving; the method is organized around the notions that (a) individuals select what they consider to be their best strategy; (b) people adapt strategies based on changing rates of success; (c) paths of development emerge as students gain experience; and (d) performance improvements are accompanied by increases in speed and a reduction in the data processed. Although each of these models contributes to a generalized framework for the development of problem-solving simulations and assessments, none alone completely satisfies the requirements outlined earlier.

Over 15 years of research in model-based, cognitively sensitive assessments (e.g., Baker, 2002, 2004, 2005) suggests that we must first focus on desired student cognition and learning, then focus on the specific subject matter (content) to develop simulations and assessments that will produce useful and usable information. Assessments that focus on student cognition and learning must address (a) content understanding; (b) problem solving; (c) metacognition; (d) communications; and (e) teamwork and collaboration. Each of these "families of cognitive demands" can be further refined. For example, content understanding can be distilled into the elements of (a) student understanding of the big ideas in a domain; (b) seeing the relationships between these big ideas; (c) avoiding misconceptions about or when using these big ideas; and (d) integrating these big ideas with prior knowledge. This framework has been used to develop performance-based assessments for the Hawaii State Assessment (Baker et al., 1996), the Los Angeles Unified School District assessment program, and the Chicago Public Schools. More germane to this volume, we have also used these models to design and make prototypes for simulation-based assessments for the U.S. Navy and to develop an online assessment design system for classroom teachers (Vendlinski et al., 2004). Others have adopted similar basic frameworks as well (e.g., Accreditation Board for Engineering and Technology, http://www.abet.org).

Typically, measurement experts have argued that accountability and diagnosis should be conducted with separate types of assessments, but for practical, economic, and conceptual reasons, we argue that they can be merged into a single measure with different methods of reporting the data for different purposes (Baker, Aschbacher, Niemi, & Sato, 1992). Findings in recent studies supported this hypothesis. So, instead of building assessments that evaluate if students have "all the facts," we are evaluating both the facts they have and how those facts are organized while realizing that the organizing principles of learners and experts are likely to be different (Chung & Baker, 2003; Doerr, 2003; Hmelo-Silver & Pfeffer, 2004; Mestre, 2000).

Building Simulations

We have used these models to construct a number of simulation information sources that are associated with a key big idea in a knowledge domain of interest. For example, working with associates at the University of Southern California, we designed a rocket ship docking simulation that requires problem solvers to dock a rocket in a number of different bays. Users can set thrust levels (amount of force) for specific amounts of time and can immediately see the resulting motion of the rocket. The simulation designer, course instructor, or the simulation itself can change the mass of the rocket either randomly for each simulated "case" or on a specified schedule, depending on pedagogical needs. The simulation

is applicable to many concepts in the knowledge domain of Newtonian force and motion, given a frictionless environment (space), but it focuses on the single organizing principle of Newton's laws in a straightforward manner. The simulation allows problem solvers an almost limitless number of solution strategies while supporting simple (manual) to complex (programmed) thrust schedules. Finally, the simulation interface records the interaction between the problem solver and the simulation for later analysis using artificial neural and Bayesian networks. The simulation allows assessors to determine student understanding of a number of concepts, ranging from accurately predicting resultant motion (vectors) to calculating and applying correct amounts of force given changes in mass or desired changes in speed and acceleration. At its heart, however, the simulation provides data to make inferences of student understanding about Newton's laws.

Based on a number of similar successes with designing assessments around information sources, we piloted an assessment template that scaffolds the integrated framework described here (Vendlinski & Niemi, 2006). The resulting system allows users the ability to intelligently design and deliver a wide range of assessments, including problem-solving simulations to students.

The Assessment Design and Delivery System Template

The Assessment Design and Delivery System (ADDS) is a powerful set of computerized tools that (a) provide utilities for individual teachers, teams of teachers, or other assessment builders to become designers and users of assessments that yield usable information to guide their pedagogy and student learning; and (b) allow designers to embed content, assessment, and pedagogical knowledge to assist teachers in both developing assessments and interpreting student progress. The ADDS is composed of four tools: the Designer, the Assembler, the Scheduler, and the Gradebook, but we only discuss the Designer in this report; the other tools are explained elsewhere (Vendlinski, Niemi, & Wang, 2005; Vendlinski, Niemi, Wang, & Monempour 2005).

The Designer acts as an assessment design template and is essential to both assessment and information source development. It instantiates the National Center for Research on Evaluation, Standards, and Student Testing models described above (see the section *Using the Characteristics to Design Simulations and Assessments for Educational Use* on page 12 of this report). Although the ADDS is useful for anyone designing an assessment, its primary intent was to infuse assessment development research directly into the classroom in a format educators found easy to use. The Designer scaffolds a teacher-user's thinking about the assessment that will be most applicable in a particular situation. Although not specifically

designed for simulation implementation, the ADDS allows designers to focus on and specify the attributes of simulation design that will make the resulting simulations both useful and usable as assessments. Scaffolding, in the form of a development template, serves both to focus the user on the essential attributes of high-quality assessment and as an aid in searching for exiting assessments. Some of the assessment attributes designers must consider are commonplace. For example, it is essential that the grade-level and linguistic complexity of the assessment item match the general ability level of the target population. This is just as true for information sources that provide the context for the assessment (such as simulations) as it is for the assessment questions that prompt the student response. Even though information sources can be useful in a number of contexts, the question prompt is designed to elicit specific student responses or to focus student attention. Nevertheless, the two must work together. The ADDS asks assessment designers to specify these attributes at the beginning of the assessment design process.

It is, however, the consideration of more atypical attributes of an assessment and the information source that the research cited here and our experiences suggest are key to developing a teacher-user's assessment acumen. Again, this is just as true in designing an information source like a simulation as it is for assessment design in general. For example, one of the most critical attributes in developing a good simulation and a good assessment is the need to specify the depth and type of knowledge a student will need to complete a task successfully. The cognitive difficulty of recalling previously presented data differs greatly from the cognitive demands of explaining an idea or constructing a more novel solution strategy. An information source that allows students an opportunity to explore a problem space and then solve a problem or generalize a solution to a set of similar problems is much more cognitively demanding than one that merely provides formulas students can use to find an answer algorithmically. Although the ADDS accommodates both types of cognitive demands, it pushes assessment designers to distinguish them and then to design assessments to fit that need by asking designers to provide this information as an assessment attribute for each assessment and information source the designer creates. Another key requirement is specification of the standard or topic (big idea) to be assessed. Although some (e.g., Stiggins, 1994) have argued the need for assessment designers to state explicitly the standard or topic to be assessed for some time, such a requirement is only becoming ubiquitous since the No Child Left Behind Act of 2001 (2002).

Given the importance of the relationship between an information source and a problem space, the ADDS encourages assessment designers to consider exactly how the information sources (such as simulations) will support the inferences about student learning they wish to

make from the assessment that is under design. Information sources can be textual, images, animation/video/audio files, or simulations. The design process for the simulations used in the ADDS follows the process used for any other information source and assessment. Initially, the grade level of the student-user is considered, followed by a determination of the domain or topic (or content standard) of interest. It should be noted that, although we have focused on state educational standards here, what is important is that there be a pedagogical goal driving the design and use of the simulation whether it is set by the state, educator, designer, or student-user, and that this goal be explicitly stated rather than remaining an implicit or ill-defined notion (Gearhart et al., 2006). Next, the designer or design team must consider the cognitive demands that the simulation will place on the student. In the ADDS, simulations could place varying demands on students. At certain times, the simulation might require students to supply recalled information; at other times, students might be asked to predict the outcomes of activities.

By asking assessment designers to supply key attributes of an assessment item or information source, the ADDS template scaffolds design and focus development on the objective of the assessment from the inception of an assessment. We use the simulation of Newtonian ramps as an example of how this process works in actual practice.

In California, middle school science students explore key concepts of Newtonian mechanics such as the relationship between unbalanced forces and motion and the relationship between force and acceleration. They also are required to understand velocity, know how to find average speed, and be conversant with graphs of both position and speed versus time. These standards, then, form the general objective of an assessment, provide an idea of the developmental level of the student, and help a test developer focus on the cognitive demands an assessment will require. In thinking about what would be required of students in this particular case, the test developer wanted various assessments that would allow the students to experiment, hypothesize, and make conclusions about force, mass, and acceleration in the Newtonian frame (the big idea). The developer also wanted a context that would allow students to collect data, present the data graphically, and interpret the data to find position and average speed at various times in the experiment.

The variety of these needs suggested that a problem-solving simulation would provide an appropriate assessment context. The test designer felt that the simulation information source could be useful for students from 4th to 10th grade because students in all those grades should find it easy to interact with the simulation, and standards in each grade deal with topics addressed by such a simulation. This is not to imply that each grade deals with force and motion. The standards in Grades 8 and 10 deal with force and motion, but the

standards in Grade 4, for example, deal with constructing graphs from measurement. We find that encouraging test developers to think about where information sources might be useful encourages them to think more about how big ideas develop over time and are interconnected rather than just testing isolated facts at a single point in students' educational careers. In fact, we have found that, unlike their peers, teachers who design assessments using the ADDS are much more likely to begin the assessment development process by noting the broad idea that they were trying to assess, and their assessments were more likely to have the students address these big ideas rather than merely recalling specific facts from a particular unit of study (Vendlinski, Niemi, & Wang, 2005).

Although we have used simulations with embedded assessment questions, we have found that embedding questions directly in the simulation (or any information source), rather than posing them outside the simulation, can limit the simulation's adaptability as a learning-and-assessment instrument and can often dramatically change the solution strategies of students. In the case of the Newtonian ramps simulation, for example, embedded assessment items not only focused the simulation on a specific standard but also tended to make the assessment less adaptable to various assessment needs (the student was prompted to find the solution anticipated by a specific question rather than explore and explain the student's understanding of the problem space). Azevedo and colleagues (Azevedo, Cromley, & Seibert, 2004), however, sought to foster student metacognition and improve learning through questions generated in response to student interactions with the system (adaptive scaffolding). They found that this type of scaffolding not only improved important student cognitive processes (planning to use and activating prior knowledge, monitoring the progress of their solution, and using multiple appropriate problem-solving strategies) more than fixed scaffolding (hard-coded questions or prompts), but also students exposed to adaptive scaffolding learned more declarative knowledge than did their counterparts using a system of fixed scaffolds. Even more surprising, the no-scaffolding condition was more effective than fixed scaffolding in promoting student metacognitive development (Azevedo, Cromley, Winters, Moos, & Greene, 2005). Moreno and Mayer (2005) also investigated adaptive scaffolding and discovered greater learning gains when they asked students to justify correct answers but not incorrect choices.

Obviously, designing simulations and assessments in this way also meant that our analytical methods needed to adapt to the nature of the assessment, and that all simulations were not equally useful in supporting all types of inferences. For example, although the clickstream data from simulations allowed us to look at the problem-solving strategies used by the students, assessments that required explanation provided more detail on student

understanding of particular facts or concepts. Again, the learning and instructional goals of the teacher and the reason for the assessment should determine the type of scaffolding or questioning present within the simulation and the analytical methods employed to make valid inferences. The assessment methods must align with educational goals and objectives (Baker, 2005; Gearhart et al., 2006).

The ADDS template guides thinking about the development of information sources and assessments, but it does not mandate that designers or developers supply every attribute of the assessment. However, what seems to be most critical is not that designers check the attribute boxes in the ADDS when designing a particular assessment item, but that designers consider these attributes when they are selecting or creating information sources (such as simulations) and designing assessments. Moreover, as the assessment or information source is used for more varied populations, to cover other topics or standards, or to assess different cognitive demands, attributes can be changed or new attributes added. As an assessment is used, details of how the assessment and the information source function in practice can be recorded in the Notes section of each ADDS assessment item. In this way, the template serves not only to scaffold the development and selection of information sources and assessments, but also to allow developers and users to organize the information sources and assessments so they can be reused from year to year or across classrooms, schools, districts, and states.

Another aspect of information source selection and assessment design we have found critical to the development of good assessment is that teachers specify or select a scoring criterion or method and have some expectation of likely student responses. This is roughly akin to a task analysis in ECD (Mislevy et al., 2004). Our experience suggests that the very process of developing these criteria encourages test writers to clarify or refine the test question. Rubrics can also be aids for instructional development and content building for teachers because teachers can now clearly see not only what their students are expected to know, but also how they will be expected to use that knowledge. Unfortunately, most teachers seldom keep their rubrics from one year to the next, and so the possibility of long-term assessment "polishing" is lost (National Research Council [NRC], 1999). In our studies with the ADDS system, we found that the system encouraged teachers to construct rubrics, and that when working without the scaffolding supplied by ADDS, teachers seldom did that on their own (Vendlinski, Niemi, & Wang, 2005).

Objects

Teachers need to have not only a deep understanding of the content knowledge and skills they are teaching and knowledge of how students develop that knowledge but also usable materials and strategies for diagnosing student learning and modifying the course of instruction when students are having difficulty (Ball, Lubienshi, & Mewborn, 2001). To address this need, we plan to provide the ADDS with an ability to make "intelligent" assessment recommendations. We will embed two objects in the ADDS to facilitate such a capability. We have borrowed the term object from computer science. In that field, an object is seen as a data structure and methods to operate on the data contained within that structure. In most cases, objects can be used by any program that complies with their input and output specifications (reusable); can inherit common data structures and methods for processing data from one or more other related objects (inheritance); accomplish a task in a way concealed from the user (encapsulation); and have clearly stated input parameters, output parameters, and ways of interacting with other objects (interaction). Learning objects are similar in that they can be reused in different contexts, interact with different applications, be used for different purposes, and have defined input and output parameters, such as a common data model associated with competency definitions, hierarchies, and maps (e.g., see http://ieeeltsc.org); however, the objects we describe for the ADDS are more properly seen as imbuing the ADDS with the ability to deliver targeted assessment objects. The first object described next manages student responses to tracked assessment items (the input), computes various psychometric parameters based on this input (a method), and passes those data to another object (interaction). This second object then uses its own data to suggest that a student has either mastered a topic or should be delivered another specific assessment item (output). Each of these objects is described in more detail next.

The first object in the sequence is a "gradebook" object. In a sense, this object is similar to Mislevy's (Mislevy, Steinberg, & Almond, 2003) student model. As is suggested by its name, this object keeps a record of available student demographic data, the tasks each student has completed, and an evaluation of how a student did on each of those tasks. This data structure currently exists in the ADDS. We plan to add methods to this object that will give it the ability to update performance records and to report various characteristics of student ability (including Item Response Theory [IRT] estimates of student ability). The object will also have methods to apply the analytical power of artificial neural networks and lag sequential analysis to classify the strategies that a student uses to solve specific types of problems.

The second object that will be added to the ADDS is an ontological object. This object builds on the ECD (Mislevy et al., 2004), RSM (Tatsuoka et al., 2004), and strategic (Stevens & Palacio-Cayetano, 2003) models by situating each problem-solving and assessment task in the context of a domain of knowledge. Based on our work and the work of our colleagues (Chung & Baker, 2003), we have developed mappings (called *ontologies*) of the relationships between big ideas within domains of knowledge as organized by experts. Each assessment task is then mapped onto this ontology based on how well the task predicts that a student understands the given concept. Although the assessment developer can make this mapping manually, we have exploited the power of Bayesian network analysis to determine the relationship between a task and a concept.

In a manner similar to that described by Mislevy et al. (2003), the ontological object uses historical data about the task and students who have accomplished the task to determine the probability that a subject who correctly or incorrectly accomplishes the task understands the concept. By using this method, the ontological object can quickly update the relationship between every assessment task and every concept in the ontology. By communicating with one another and the ADDS template interface, the two objects are able to isolate the concepts an educator wants to assess, the ability of a student or of a student group given responses to other tasks, and the likelihood that the individual or group understands that particular concept. Based on this analysis, the ADDS can then recommend the most appropriate next assessment task to the teacher.

As currently envisioned, the system will recommend the task that most dramatically improves our estimate that a student (or student group) understands a specific concept.[3] Because the Bayesian net calculates the probability of understanding a concept based both on the results of individual tasks and the likelihood that other concepts are understood, the system may recommend assessing a different, more fundamental concept than initially identified by the teacher. We suspect that more accurate inferences of student understanding and ability might be possible by exploiting a number of different indicators (item technical parameters such as reliability and generalizability measures [Shavelson & Webb, 1991]; neural network analysis [Principe, Euliano, & Lefebvre, 2000]; Markov models [Rabiner, 1989]; and lag sequential analysis [Bakeman & Gottman, 1997]).

[3] For each task associated with a concept, the ontological object will calculate the difference between the probability that a student understands the concept given successful task completion and the probability that a student understands the concept given unsuccessful task completion. The object then returns the task with the maximum difference.

Summary and Discussion

Our experience developing problem-solving simulations and assessments suggests that the following must be specified at the start of the assessment or simulation design process and be considered throughout that process:

- The simulation or assessment has a specific goal or learning outcome. In the ADDS template, this goal is represented by a state standard and an ontological representation of how knowledge (a big idea) develops in the domain of interest.
- The cognitive level required to complete the task (or each subtask) successfully is clearly stated, and the problem allows the differentiation of levels of understanding. Problem-solving simulations in the ADDS template can have multiple levels of cognitive demand (recall, explain, problem-solving application, make connections to other knowledge, transfer) depending on the questions that wrap the simulation. We have created simulations in which a user could solve the simulation by merely recalling the oxidation states of two elements or memorize what occurs next in an algorithmic problem-solving context. The user's ability to do this was closely linked to limited types of problem-solving strategies. Ultimately, then, "whatever task a teacher poses, its cognitive demand is shaped by the way [they have] students use it" (Kilpatrick, Swafford, & Findell, 2001, p. 335). As we have suggested, one can make tasks more cognitively demanding by varying the context (information source) and assessment demands (questions prompts) of a task, and these variations seem essential not only to ameliorating misconceptions, but also to preventing them from forming in the first place.
- The problem has associated criteria or methods that an evaluator can use to accurately evaluate the problem-solving performances generated by users. Our work and the work of colleagues repeatedly concluded that specifying criteria as part of the development process is correlated with improved assessment tasks. For the Newtonian ramp problem, there was a single correct solution but no single way to arrive at the solution, and a user's ability to solve such a problem seemed more tied to deep understanding than mere recall of fact. Detailed scoring criteria provide users a metric to discern various levels of performance. Moreover, when educators consider the evaluation criteria associated with an assessment task as part of their choice of an assessment, instruction is improved. It must also be remembered that research suggests that a mix of various assessment formats (selected response, explanation, application, computer based, pencil and paper, real world, or "wet labs") provides the most accurate indication of student ability.
- The complexity of the assessment or simulation should correspond to the learning and instructional goals of the educator or course of instruction.
- Finally, users must consider the impact of external constraints. Timed conditions, connectivity, and hardware (including variables such as the speed of connectivity to the World Wide Web, the computer [and network] operating system [especially Mac/PC], available memory, and support software) can have detrimental effects on the usefulness of computer-based assessments.

References

Anderson, K. C., & Leinhardt, G. (2002). Maps as representations: Expert novice comparison of projection understanding. *Cognition and Instruction, 20,* 283–321.

Azevedo, R., Cromley, J. G., & Seibert, D. (2004). Does adaptive scaffolding facilitate students' ability to regulate their learning with hypermedia? *Contemporary Educational Psychology, 29,* 344–370.

Azevedo, R., Cromley, J. G., Winters, F. I., Moos, D. C., & Greene, J. A. (2005). Adaptive human scaffolding facilitates adolescents' self-regulated learning with hypermedia. *Instructional Science, 33,* 381–412.

Bakeman, R., & Gottman, J. M. (1997). *Observing interaction: An introduction to sequential analysis*. Cambridge, UK: Cambridge University Press.

Baker, E. L. (1998). *Model-based performance assessment* (CRESST Tech. Rep. No. 465). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Baker, E. L. (2002). Design of automated authoring systems for tests. In N. R. Council (Ed.), *Technology and assessment: Thinking ahead* (pp. 79–89). Washington, DC: National Academy Press.

Baker, E. L. (2004, March). *Assessing and monitoring performance across time and place*. Paper presented at the U.S. Department of Education Secretary's No Child Left Behind Leadership Summits "Empowering Accountability and Assessment Using Technology," St. Louis, MO.

Baker, E. L. (2005). Technology and effective assessment systems. In J. L. Herman & E. H. Haertel (Eds.), *Uses and misuses of data for educational accountability and improvement. NSSE Yearbook* (Vol. 104, pp. 358–378). Chicago: National Society for the Study of Education.

Baker, E. L., Aschbacher, P. R., Niemi, D., & Sato, E. (1992). *CRESST performance assessment models: Assessing content area explanations* (CRESST Tech. Rep. No. 652). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Baker, E. L., Freeman, M., & Clayton, S. (1991). Cognitive assessment of history for large-scale testing. In M. C. Wittrock & E. L. Baker (Eds.), *Testing and cognition* (pp. 131–153). Englewood Cliffs, NJ: Prentice-Hall.

Baker, E. L., Niemi, D., Herl, H., Aguirre-Muñoz, Z., Staley, L., & Linn, R. L. (1996). *Report on the content area performance assessments (CAPA): A collaboration among the Hawaii Department of Education, the Center for Research on Evaluation, Standards, and Student Testing (CRESST) and the teachers and children of Hawaii* (Final Deliverable). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Ball, D. L., Lubienshi, S. T., & Mewborn, D. S. (2001). Research on teaching mathematics: The unsolved problem of teachers mathematical knowledge. In V. Richardson (Ed.), *Handbook of research on teaching* (4th ed., pp. 433–456). Washington, DC: American Educational Research Association.

Black, P., & Wiliam, D. (2004). The formative purpose: Assessment must first promote learning. In M. Wilson (Ed.), *Towards coherence between classroom assessment and accountability* (Vol. 2, pp. 20–50). Chicago: University of Chicago Press.

Bobick, A. F. (1987). *Natural object categorization*. Unpublished doctoral dissertation, Massachusetts Institute of Technology, Cambridge.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (Eds.). (1999). *How people learn: Brain, mind, experience, and school*. Washington, DC: National Academy Press.

Bransford, J. D., & Schwartz, D. L. (Eds.). (1999). *Rethinking transfer: A simple proposal with multiple implications* (Vol. 24). Washington, DC: American Educational Research Association.

Buckley, B. C., & Boulter, C. J. (2000). Investigating the role of representations and expressed models in building mental models. In J. K. Gilbert & C. J. Boulter (Eds.), *Developing models in science education* (pp. 105–122). Dordrecht, The Netherlands: Kluwer.

Buckley, B. C., Gobert, J. D., Gerlits, B., Goldberg, A., & Swiniarski, M. J. (2004, April). *Assessing model-based learning in BiloLogica*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Caine, R. N., Caine, G., McClintic, C., & Klimek, K. (2005). *Twelve brain/mind learning principles in action: The fieldbook for making connections, teaching, and the human brain*. Thousand Oaks, CA: Corwin Press.

Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5*, 121–152.

Chung, G. K. W. K., & Baker, E. L. (2003). An exploratory study to examine the feasibility of measuring problem-solving processes using a click-through interface. *Journal of Technology, Learning, and Assessment, 2*(2), 1–30.

Chung, G. K. W. K., Harmon, T. C., & Baker, E. L. (2001). The impact of a simulation-based learning design project on student learning. *IEEE Transactions on Education, 44*, 390–398.

Cromley, J. G., & Mislevy, R. J. (2004). *Task templates based on misconception research* (CRESST Tech. Rep. No. 646). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

DeCorte, D., Greer, B., & Vershaffel, L. (1996). Mathematics teaching and learning. In D. Berliner & R. Calfee (Eds.), *Handbook of educational psychology* (pp. 491–549). New York: Simon and Schuster Macmillan.

Doerr, M. (2003). The CIDOC conceptual reference module: An ontological approach to semantic interoperability of metadata. *AI Magazine, 24*, 75–92.

Dweck, C. S. (2002). The development of ability conceptions. In A. Wigfield & J. Eccles (Eds.), *The development of achievement motivation* (pp. 57–88). New York: Academic Press.

Edelstein, R. A., Reid, H. M., Usatine, R., & Wilkes, M. S. (2000). A comparative study of measures to evaluate medical students' performances. *Academic Medicine, 75*, 825–833.

Ericsson, K. A. (2003). The search for general abilities and basic capacities: Theoretical implications from the modifiability and complexity of mechanisms mediating expert performance. In R. J. Sternberg & E. L. Grigorenko (Eds.), *Perspectives on the psychology of abilities, competencies, and expertise* (pp. 93–125). Cambridge, UK: Cambridge University Press.

Fuchs, L. S., Fuchs, D., Prentice, K., Hamlett, C. L., Finelli, R., & Courey, S. J. (2004). Enhancing mathematical problem solving among third-grade students with schema-based instruction. *Journal of Educational Psychology, 96*, 635–647.

Gearhart, M., Nagashima, S., Pfotenhauer, J., Clark, S., Schwab, C., Vendlinski, T., et al. (2006). Developing expertise with classroom assessment in K–12 science: Learning to interpret student work interim findings from a 2-year study. *Educational Assessment, 11*, 237-263.

Glaser, R. (2001). Conflicts, engagements, skirmishes, and attempts at peace. *Educational Assessment, 7*, 13–20.

Glaser, R., & Baxter, G. P. (2000). *Assessing active knowledge* (CRESST Tech. Rep. No. 516). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Gobert, J., Buckley, B., & Clarke, J. E. (2004, April). *Scaffolding model-based reasoning: Representation, cognitive affordances, and learning outcomes.* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Gould, S. J. (1996). *The mismeasure of man.* New York: Norton.

Gredler, M. E. (2004). Games and simulations and their relationships to learning. In D. H. Jonassen (Ed.), *Handbook of research on educational communications and technology* (pp. 571–582). Mahwah, NJ: Lawrence Erlbaum Associates.

Greer, B. (1992). Multiplication and division as models of situations. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 276–295). New York: Macmillan.

Halabi, A. K., Tuovinen, J. E., & Farley, A. (2005). Empirical evidence on the relative efficiency of worked examples versus problem-solving exercises in accounting principles instruction. *Issues in Accounting Education, 20*, 21–32.

Herman, J. (1992). *Accountability and alternative assessment: research and development issues* (CRESST Tech. Rep. No. 348). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Hmelo-Silver, C. E., & Pfeffer, M. G. (2004). Comparing expert and novice understanding of a complex system from the perspective of structures, behaviors, and functions. *Cognitive Science, 28*, 127–138.

Jacobson, M. J. (2001). Problem solving, cognition, and complex systems: Differences between experts and novices. *Complexity, 6*(2), 1–9.

Kilpatrick, J., Swafford, J., & Findell, B. (Eds.). (2001). *Adding it up: Helping children learn mathematics.* Washington, DC: National Academy Press.

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist, 41*, 275–286.

Leemkuil, H., Jong, T. D., & Ootes, S. (2000). *Review of educational use of games and simulations*. Ae Enschede, The Netherlands: University of Twente.

Mayer, R. E. (2003). *Learning and instruction*. Upper Saddle River, NJ: Merrill Prentice-Hall.

Mayer, R. E., & Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educational Psychologist, 38*, 43–52.

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

Mestre, J. P. (2000). Progress in research: The interplay among theory, research questions and measurement techniques. In A. E. Kelly & R. A. Lesh (Eds.), *Handbook of research design in mathematics and science education* (pp. 151– 168). Mahwah, NJ: Lawrence Erlbaum Associates.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2004). *A brief introduction to evidence-centered design* (CRESST Tech. Rep. No. 632). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*, 3–66.

Mislevy, R. J., Steinberg, L. S., Breyer, F. J., Almond, R. G., & Johnson, L. (1999). A cognitive task analysis with implications for designing simulation based performance assessment. *Computers in Human Behavior, 15*, 335–374.

Moreno, R., & Mayer, R. (2005). Role of guidance, reflection, and interactivity in an agent-based multimedia game. *Journal of Educational Psychology, 97*, 117–128.

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics: An overview*. Washington, DC: Author.

National Research Council. (1999). *Global perspectives for local action: Using TIMSS to improve U.S. mathematics and science education*. Washington, DC: National Academies Press.

Newell, A., & Simon, H. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

No Child Left Behind Act of 2001, 20 U.S.C. § 6301 (2002).

Norman, D. A. (1993). *Things that make us smart: Defending human attributes in the age of the machine*. New York: Addison-Wesley.

Paas, F., Renkl, A., & Sweller, J. (2003). Cognitive load theory and instructional design: Recent developments. *Educational Psychologist, 38*, 1–4.

Pellegrino, J., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.

Principe, J. C., Euliano, N. R., & Lefebvre, W. C. (2000). *Neural and adaptive systems*. New York: Wiley.

Quellmalz, E. S., & Hartel, G. (2004). *Technology supports for state science assessment systems*. Washington, DC: National Research Committee on Test Design for K–12 Science Achievement.

Quinn, C. N. (2005). *Engaging learning: Designing e-learning simulation games*. San Francisco: Jossey-Bass.

Rabiner, L. (1989). A tutorial on hidden Markov models and selected application in speech recognition. *Proceedings of the IEEE, 77*, 257–286.

Randel, J. M., Morris, B. A., Wetzel, C. D., & Whitehill., B. V. (1992). The effectiveness of games for educational purposes: A review of recent research. *Simulation and Gaming, 23*, 261–276.

Renkl, A. (2002). Worked-out examples: Instructional explanations support learning by self-explanation. *Learning and Instruction, 12*, 529–556.

Richards, W., Feldman, J., & Jepson, A. (1992). From features to perceptual categories. In D. Hogg & R. Boyle (Eds.), *British Machine Vision Conference 1992* (pp. 99–108). Berlin: Springer-Verlag.

Rieber, L. P. (2005). Multimedia learning in games, simulations, and microworlds. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 549–567). Cambridge, UK: Cambridge University Press.

Rogoff, B. (1998). Cognition as a collaborative process. In D. Kuhn & R. S. Siegler (Eds.), *Handbook of child psychology: Cognition, perception and language* (Vol. 2, pp. 679–744). New York: Wiley.

Schraw, G., Brooks, D., & Crippen, K. J. (2005). Using an interactive, compensatory model of learning to improve chemistry teaching. *Journal of Chemical Education, 82*(4), 637–640.

Schworm, S., & Renkl, A. (2006). Computer-supported example-based learning: When instructional explanations reduce self-explanations. *Computers and Education, 46*, 426–445.

Seel, N. M., & Schenk, K. (2003). An evaluation report of multimedia environments as cognitive learning tools. *Evaluation and Program Planning, 26*, 215–224.

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

Steffe, L. P., & Thompson, P. W. (2000). Teaching experiment methodology: Underlying principles and essential elements. In A. E. Kelly & R. A. Lesh (Eds.), *Handbook of research design in mathematics and science education* (pp. 267–306). Mahwah, NJ: Lawrence Erlbaum Associates.

Stevens, R., & Palacio-Cayetano, J. (2003). Design and performance frameworks for constructing problem-solving simulations. *Cell Biology Education, 2*, 162–179.

Stiggins, R. J. (1994). *Student-centered classroom assessment*. New York: Macmillan College.

Sweller, J. (2003). Evolution of human cognitive architecture. In B. Ross (Ed.), *The psychology of learning and motivation* (Vol. 43, pp. 215–266). San Diego, CA: Academic Press.

Sweller, J., & Cooper, G. A. (1985). The use of worked examples as a substitute for problem solving in learning algebra. *Cognition and Instruction, 2*, 59–89.

Tatsuoka, K. K., Corter, J. E., & Tatsuoka, C. (2004). Patterns of diagnosed mathematical content and process skills in TIMMS-R across a sample of 20 countries. *American Educational Research Journal, 41*, 901–926.

VanLehn, K. (1990). *Mind bugs*. Cambridge, MA: MIT Press.

Vendlinski, T. P. (2001). *Affecting U.S. education through assessment: New tools to discover student understanding*. Unpublished doctoral dissertation, Massachusetts Institute of Technology, Cambridge.

Vendlinski, T. P., Baker, E. L., & Niemi, D. (2008). Templates and objects in authoring problem-solving assessments. In E. Baker, J. Dickieson, W. Wulfeck & H. F. O'Neil (Eds.), *Assessments of problem solving using simulations* (pp. 309–333). New York: Lawrence Erlbaum Associates.

Vendlinski, T. P., Munro, A., Bewley, W. L., Chung, G. K. W. K., Pizzini, Q., Stuart, G., et al. (2004, December 6–9). *Proceedings of the Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC) 2004. Learning complex cognitive skills with an interactive job aid*. Orlando, FL: National Training Systems Association.

Vendlinski, T. P., & Niemi, D. (2006, April). *Making simulations educationally beneficial*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Vendlinski, T., Niemi, D., & Wang, J. (2005, March). *Learning assessment by designing assessments*. Paper presented at the Society for Information Technology and Teacher Education (SITE) 16th International Conference, Phoenix, AZ.

Vendlinski, T., Niemi, D., Wang, J., & Monempour, S. (2005, July). *Improving formative assessment practice with educational information technology*. Paper presented at the Third International Conference on Education and Information Systems, Technologies and Applications (EISTA 2005), Orlando, FL.

Vendlinski, T., & Stevens, R. (2000). The use of artificial neural nets (ANN) to help evaluate student problem solving strategies. In B. Fishman & S. O'Connor-Divelbiss (Eds.), *International Conference of the Learning Sciences: Facing the Challenges of Complex Real-World Settings* (pp. 108–114). Mahwah, NJ: Lawrence Erlbaum Associates.

Vygotsky, L. S. (1962). *Thought and language* (C. M. V. John-Steiner, S. Scribner & E. Souberman, Trans.). Cambridge, MA: MIT Press.

Zull, J. (2002). *The art of changing the brain: Enriching the practice of teaching by exploring the biology of learning*. Sterling, VA: Stylus.