CRESST REPORT 800

Robert J. Mislevy

EVIDENCE-CENTERED DESIGN FOR SIMULATION-BASED ASSESSMENT

JULY, 2011

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

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

Evidence-Centered Design for Simulation-Based Assessment

CRESST Report 800

Robert J. Mislevy
Educational Testing Service

July, 2011

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 here was supported in part by the Center for Advance Technology in Schools (CATS), PR/Award Number R305C080015, as administered by the Institute of Education Sciences, U.S. Department of Education.

To cite from this report, please use the following as your APA reference: Mislevy, R.J. (2011). *Evidence-centered design for simulation-based assessment.* (CRESST Report 800). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

TABLE OF CONTENTS

EVIDENCE CENTERED-DESIGN FOR SIMULATION-BASED ASSESSMENT[1]

Robert J. Mislevy
Educational Testing Service

Abstract

Simulations provide opportunities for individuals to learn and develop skills for situations that may be expensive, time-consuming, or dangerous. Careful design can support a person's learning by tailoring the features of situations to his or her skill level(s), allowing repeated attempts, and by providing timely feedback. Furthermore, simulations can emulate certain environments, which in turn, provide opportunities for assessing people's capabilities to act in these situations. This report describes an assessment design framework that can help projects develop effective simulation-based assessments. The report reviews the rationale and terminology of the 'evidence centered' assessment design (ECD) framework (Mislevy, Steinberg, & Almond, 2003), discusses how the framework aligns with the principles of simulation design, and illustrates the ideas it presents with examples from the fields of engineering and medicine. Advice is offered for designing a new simulation-based assessment and for adapting an existing simulation system for assessment purposes.

Introduction

Advances in technology open the door to a radically new paradigm of learning, which is characterized by the interaction and adaptation that simulation environments afford. Simulations allow people to explore phenomena that may be excessively fast, slow, expensive, time-consuming, or dangerous (Snir, Smith, & Grosslight, 1993). With physical simulators, for instance, physicians can practice heart surgery and pilots can land planes with failed engines. Moreover, in interactive digital simulations, students can work with groups of stakeholders to revitalize a town center as urban planners; in the field of medicine, students can practice diagnosis and patient care. What is more, in role playing scenarios, professionals can have face-to-face interactions with real people in situations designed to help them develop experience with conflict resolution or emotionally difficult situations (e.g., medical patients).

Given the increased acceptance and widespread use of simulations for learning, it is natural to consider the utility of simulation environments for assessment purposes. In fact, several uses of simulations are currently online—including the computer-based site design

[1]The author is grateful to many colleagues (from Cisco, National Center for Research on Evaluation, Standards, and Student Testing, The Dental Interactive Simulations Corporation, Educational Testing Service, SRI International, and the University of Maryland) for their insights and collaboration.

problems in the National Council of Architectural Registration Boards' Architectural Registration Examination (ARE) (Bejar & Braun, 1999), as well as the simulated-patient encounters and the Primum computer-based patient management problems in Step 3 of the United States Medical Licensing Examination (Dillon, Boulet, Hawkins, & Swanson, 2004).

However, the move from simulation to simulation-based assessment is not simple. For instance, the principles and tools needed to create valid assessment in simulation environments are not the same as those required to build simulations (or even to use them for learning) (Melnick, 1996). One challenge is that the development of a valid simulation-based assessment requires that expertise from disparate domains come together to serve the assessment's purpose (typically including subject matter knowledge, software design, psychometrics, assessment design, and pedagogical knowledge). Few people are experts in all of these domains; fortunately, it is sufficient to have a shared design framework in which each team member can see how his expertise fits in with others.

This report will describe such a framework—namely that of evidence-centered assessment design (ECD) (Almond, Steinberg, & Mislevy, 2002; Mislevy, Steinberg, & Almond, 2003; Mislevy & Riconscente, 2006). ECD builds on recent research in assessment design to provide language as well as representations that help assessment designers across domains, task types, and purposes. ECD, for example, has been used to design such diverse assessments as Advanced Placement tests in science and history (Huff, Steinberg, & Matts, 2010), teacher certification examinations (Pearlman, 2004), and commercial vehicle driver assessments (Haertel, Wentland, Yarnall, & Mislevy, in press). Moreover, ECD has proven its usefulness as the design framework for a number of simulation-based assessments—including applications in computer network engineering (Behrens et al., in press; Williamson et al., 2004), science investigations in virtual worlds (Clarke-Midura, Code, Dede, Mayrath, & Zap, in press), and problem-solving in dental hygiene (Mislevy, et al., 2002).

This report begins with a brief discussion of relevant principles from learning psychology and simulation design. An overview of ECD will then be presented and each layer in the design process will be discussed further. Space limitations preclude detailed discussions; thus, references will be provided for different aspects of the framework as well as worked examples. Advice is then presented for two cases: 1) developing a simulation-based assessment from existing simulation capability, and 2) creating a simulation-based assessment de novo.

Background

In his review of the ways in which people become experts, Salthouse (1991) found that novices face similar difficulties across a wide variety of domains. He discovered that novices did not know what information was relevant in a situation or how to integrate pieces of information they possessed. Furthermore, they did not recognize what to expect or what to do (either as possible courses of action or ways to choose them). Surprisingly, even when novices had an idea of what to do, they often could not do it well or quickly enough.

Humans routinely display noteworthy capabilities; yet, these capabilities are so commonplace that we fail to appreciate them. For example, in milliseconds, individuals can simultaneously carry on a rapid back-and-forth conversation by processing sounds, grammar, domain knowledge, social norms, conversational conventions, and pragmatic moves. How does this occur? It happens because people are extremely good at working with patterns (e.g., patterns of perceiving, thinking, and acting); as a result, once we are sufficiently practiced, we become attuned to and can assemble patterns flexibly in real-time. Expert performance is made possible through the continual interaction between the external patterns that structure people's interactions in situations and with others (e.g., language and professional practices) as well as each person's internal neural patterns for recognizing, making meaning of, and acting through these patterns (Wertsch, 1998). We develop our internal cognitive patterns through experience by participating in activities that are structured around external patterns and by discerning the regularities by seeing what happens as others act (and when we ourselves act); as a result, we become more flexible and capable in increasingly broader ranges of situations.

More specifically, people become experts in domains such as sports, engineering, medicine, and culinary arts by spending time taking part in the activities of that domain (e.g., learning to work on the problems, read the literature, talk with the people, and act in the situations). Experts learn to use the tools and strategies that have been developed in the community (Ericsson, Charness, Feltovich, & Hoffman, 2006). Furthermore, those who become experts build their capabilities and overcome the pervasive limitations that plague novices through reflective practice (which is best accomplished with feedback and often starts from simplified situations that are usually supported by others). Although experts generally know more than novices, it is important to highlight that experts' knowledge is organized around underlying principles in the domain. Experts' knowledge is enmeshed with possible actions and ways of interacting and evolving; moreover, it is rooted in situations with people who are themselves acting and interacting (e.g., as collaborators, subjects, or adversaries) (Chi, Glaser, & Farr, 1988). Every situation is different, but experts recognize

the features and possibilities afforded by these recurring patterns (Greeno, 1998). When the assessment of knowledge and skill is discussed in this report, these are the kinds of capabilities we have in mind; this leads into a discussion of simulations.

Simulation environments highlight the key features of relevant situations—such as practice, feedback, determining what will occur, honing skills and gaining facility with tools, and building experience about what does and does not work at certain times. For instance, the first time an airline pilot's engine fails, he may have worked through a hundred similar situations in a full motion simulator. Simulation environments can enable medical students to work through months of simulated patient care in an hour, experience triage for large-scale disasters, and practice how to manipulate bronchoscopes while learning what to look for. Another example is Cisco's Packet Tracer simulation, which provides step-by-step animations of exactly what occurs in lightning fast exchanges across routers, so that students can build up mental models of what happens and why; this knowledge and these skills can be used automatically and intuitively when students work with real networks (Frezzo, 2009). All of these examples provide experiences for building up the necessary patterns to know what is important, what it means, what can be done, and how to interact with people and situations in the domain. They highlight the key patterns, allow both repetition and diverse practice, and provide critical opportunities for feedback.

It is important to note that simulations are not identical to real-life situations; hence, a simulator may emulate some features of real conditions but not others. Simulations may speed up, slow down, change size, or simplify aspects of real-world situations; moreover, simulations make it possible to replicate or vary situations in systematic ways. Within the limits of practicality, it is up to the designer to make choices about which aspects of real situations are emulated in a simulation environment, the ways and extent to which real situations are emulated, and which aspects are modified and/or omitted. Decisions regarding how simulations are to be created cannot be made without specifying the intended purposes of the simulation. In particular, higher fidelity to real-world situations does not necessarily make for better learning or better assessment. Fidelity with respect to targeted knowledge or skill (at the targeted level) is more important; yet, even that is not the whole story.

Designing simulations for *learning* requires focusing on the features of situations that provoke the targeted knowledge and skills, at a level that is just beyond the capabilities of the student, which also known in Vygotsky's (1978) terminology, as the zone of proximal development (ZPD). Given the learners' current level, we must ask: What are aspects of situations that will best solidify skills, add the next layer of complexity, or add experience with variations of a new theme? The features of situations in the domain that are critical for

the student to interact with, the means for the student to act on the system (its affordances) in order to express choices, and the reactions of the system to the student's actions (that reflect the underlying principles of the system) must all be included in the simulation.

What should be minimized are features that require too much knowledge or skills that are not central for the aspects of capability we care about or that add irrelevant complexity (Chandler & Sweller, 1991, call this "extraneous cognitive load"). A system that models the way experts view problems with high fidelity may not be accessible to beginning students (who may not know how to begin making sense of this model) (Roschelle, 1997). It proves useful to leave some components of real-world situations out of simulated situations in order to begin at a place that makes sense to beginning level examinees. Complexity can be introduced in stages to maximize effective learning. Similarly, providing opportunities to slow down or to stop action to reflect on what is important in a situation, what to do next, or why something has just happened can helps students build knowledge structures. Providing students with opportunities to wrestle with a problem in-depth, to repeat all or parts of a problem, to test hypotheses and lastly, to explore alternative options are additional ways that a well- targeted simulation environment can take advantage of the principles of learning.

Designing simulations for *assessment* requires focusing on the information about the knowledge and skill that the user needs. Sometimes the user and the examinee are the same person, (e.g., when simulations are used for self-assessment or as coached practiced systems) (Shute & Psotka, 1996). In these cases, the simulation situations are again tuned for learning but have additional processes that provide feedback about performance. Although in most cases, the users are different from the examinee; hence, the focus is usually to acquire information to evaluate examinees' capabilities, either as to overall proficiency or more specific aspects of knowledge and skill. The goal may be low-stakes formative feedback or even for higher-stakes use (such as licensure or hiring decisions).

Much of the rationale in designing a simulation for learning overlaps with designing a simulation-based assessment because the kinds of problems and situations in which one learns to think and act in a domain are the same kinds of situations in which evidence about that knowledge and skill can be evoked for assessment purposes. Exactly how that evidence is to be evoked, captured, interpreted, summarized, and reported is the aegis of assessment designers and psychometricians. An assessment design framework coordinates assessment designers and psychometricians' contributions with those of the domain experts and users.

An Overview of Evidence-Centered Design

An educational assessment is an evidentiary argument for reasoning what students say, do, or make in particular task situations as well as to generally claim what they can know, do, or have accomplished. Messick (1994) succinctly summarizes its form:

> [We] would begin by asking what complex of knowledge, skills, or other attributes should be assessed, presumably because they are tied to explicit or implicit objectives of instruction or are otherwise valued by society. Next, what behaviors or performances should reveal those constructs, and what tasks or situations should elicit those behaviors? Thus, the nature of the construct guides the selection or construction of relevant tasks as well as the rational development of construct-based scoring criteria and rubrics. (p. 16).

The evidence-centered assessment design (ECD) framework distinguishes layers at which activities and structures in assessment work to instantiate an assessment argument in operational processes. Common language and representations—across different forms of assessment—helps designers provide guidance across design, analysis, deployment, and measurement aspects. An evidence-centered assessment design (ECD) framework can help designers structure their work, both conceptually and operationally, in ways that encourage reusability (e.g., design patterns for generating tasks, adaptable scoring procedures).

Figure 1 summarizes the ECD layers in a manner that reflects successive refinement and organization of knowledge of the content domain and the purpose of the assessment—from its substantive argument to the specific elements and processes that are needed in its operation. The figure may suggest a sequential design process, but every assessment developer or simulation designer knows that work proceeds in cycles of iteration and refinement between and within layers.

The first layer, *Domain analysis* refers to marshaling substantive information about the domain. It helps us understand the kinds of problems and situations people deal with, the knowledge and skills they draw upon, the representational forms they use, and the characteristics of good work. This is the substantive foundation for an assessment.

In the *Domain Modeling* layer, information identified in Domain Analysis is organized along the lines of assessment arguments. Supporting tools such as the design patterns discussed below help developers think through the assessment argument without getting tangled up in the details of implementation. Generative schemas for families of tasks are important for assessments that need to generate multiple forms.

The *Conceptual Assessment Framework (CAF)* concerns technical specifications for operational elements. An assessment argument is now expressed in terms of coordinated

pieces of machinery such as measurement models, scoring methods, and delivery requirements. The data structures and reusability of the CAF models help bring down the costs of task design.

The fourth layer, the *Assessment Implementation* encompasses (possibly ongoing) activities that prepare for operational administration (such as authoring tasks, calibrating psychometric models, piloting and finalizing evaluation procedures, and producing assessment materials and presentation environments).

Assessment Delivery addresses the processes of presenting tasks to examinees, evaluating performances to assign scores, and reporting the results to provide feedback or support decision making. The four-process architecture is viewed in terms of processes and messages whose (a) meaning is grounded in Domain Modeling, (b) structure is laid out in the CAF, and (c) pieces are built in Assessment Implementation.

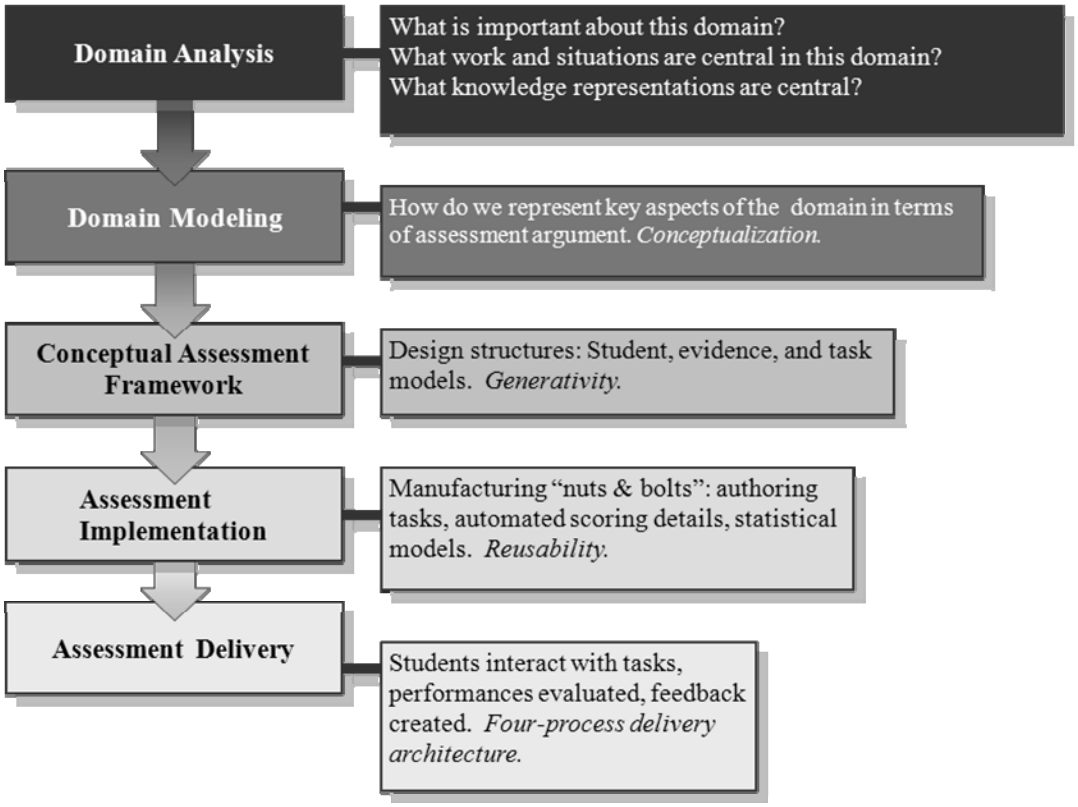

Figure 1. Layers of evidence-centered design.

Walking Through the Layers

This section takes a closer look at the ECD layers, with a special eye toward simulation-based assessment.

Domain Analysis

Domain analysis defines and documents the content or experiential domains to be assessed. In other words, domain analysis looks at the ways in which people use relevant content knowledge as well as the kinds of situations in which they would use it. For example, what constitutes mathematics, troubleshooting, or teamwork for the context at hand? What do we know about progressions of thought or skill, patterns of errors or common representations? What knowledge, skills, goals, tools, enablers, representations and possible distractions are relevant? How do people interact with the physical environment, conceptual representations, and other people in order to accomplish goals? By what standards are these efforts judged in practice? Designers can consider the domain from a number of perspectives, such as cognitive research, available curricula, professional practice, ethnographic studies, expert input, standards and current testing practices, as well as the various requirements, resources and constraints to which the proposed assessment might be subject.

Salthouse (1991) notes that for performance-based assessments, designers must understand the kinds of situations that pose recurring problems. People's patterns of thinking and acting in such situations will be at the heart of the assessment, in forms and at levels that match its purposes; hence, practitioners, researchers, and instructors' insights are invaluable.

The critical incident technique (Flanagan, 1954) has proven useful in domains such as medicine, engineering, and law enforcement. Task analysis, in which people that actually work in a domain are observed and interviewed, has long been a staple of licensure testing. More recent methods of cognitive task analysis (CTA) (Schraagen, Chipman, & Shalin, 2000) uncover the knowledge structures that people use as well as how their knowledge is related to their performance. This approach is especially suited to the developing, learning, and assessment environments—where what matters in design is not just what people do but how and why they do it. The following are relevant examples:

- HYDRIVE was a coached practice system to help Air Force trainees learn to troubleshoot the hydraulics system of the F-15 aircraft (Steinberg & Gitomer, 1996). The CTA showed that expert troubleshooting required a conjunction of knowledge of the subsystems sufficient to run 'mental models' of how they would function normally and with various faults, proficiency with the troubleshooting tools and procedures, and an understanding of strategies of problem-solving in finite domain. The last of these proved to be well described by Newell & Simon's (1972) cognitive model for problem-solving, which was subsequently used as a basis for task design, performance evaluation, and instruction.
- Katz's (1994) studies of expert and novice architects' design solutions helped ground a family of design problems for the Architectural Registration Examination. He found that the design process was invariably iterative: experts and novices alike

started from an initial solution that met some constraints, and modified it repeatedly to accommodate more, always working from the representation generated thus far. But while both experts and novices continually revised aspects of provisional designs as they progressed, the novices' rework was more often substantial and discarded much previous work. The novices encountered conflicting and hard-to-meet constraints when they were further along; whereas, the experts had identified and addressed these challenges early on. Varying the number of constraints, the challenge of meeting them, and the degree of conflict among them are systematic and cognitively relevant ways to vary task difficulty.

- Cameron et al. (1999) identified nine broad classes of behaviors that tended to distinguish along the dental hygiene expert-novice continuum: 1) Gathering and Using Information, 2) Formulating Problems and Investigating Hypotheses, 3) Communication and Language, 4) Scripting Behavior, 5) Ethics, 6) Patient Assessment, 7) Treatment Planning, 8) Treatment, and 9) Evaluation. For example, novices might identify salient features in multiple representations such as soft tissue charts and radiographs, but the experts tended to produce a conception of a patient's etiology that would lead to the pattern of features across representations.

In each of these cases, we see that results from the CTA significantly impacted thinking about the assessment argument in the project, hence, simulator design choices.

Domain Modeling

In domain modeling, designers organize information from domain analyses to describe relationships among target knowledge and skill, what we might see people say, do, or make as evidence, and situations and activities that evoke it—in short, the elements of assessment arguments. Graphical and tabular representations and schemas support this work. Among the representational forms that have been used to implement ECD are "claims and evidence" worksheets (Ewing, Packman, Hamen, & Thurber, 2010), Toulmin diagrams for assessment arguments (Mislevy, 2003), and "design patterns" for constructing assessment arguments for some aspect of capabilities (such as design under constraints and model-based reasoning) (Mislevy, Riconscente, & Rutstein, 2009). Let us look more closely at the latter two.

Figure 2 maps Messick's quote into a more formal schema based on Toulmin's (1958) schema for arguments. At the top is a claim about an examinee's knowledge and skill. At the bottom is the observation of the examinee acting in a certain situation. An assessor's reasoning moves up through principled reasoning when focusing on what is important in the examinee's actions (which the examinee produces) as well as the features of the situation that are important in provoking those actions (partly determined by the assessment designer but also partly determined by the examinee and their own interactions with the task). These two kinds of data support the claim through a warrant or rationale as to why examinees with particular knowledge or skill are likely to act in certain ways during the situation at hand.

The backing for this warrant, as well as the warrants regarding what is important about the situation and the performance, come from the Domain Analysis. This report provided examples of this grounding in the previously cited expertise studies.

The single boxes for "data concerning the student" and "data concerning the situation" are sufficient for self-contained small tasks. More needs to be said about performances (i.e., simulations that evolve over time) as the examinee interacts with them. Figure 3 shows how in such situations, some features of both the situation and the performance (e.g., a final solution) can be evaluated without attending to the details of the interaction. However other features of the situation, as well as other potentially useful data, emerge along the way; thus, the evaluation must recognize and take those situational features into account (e.g., did the examinee in a role play work herself into an apology situation; if so, was an apology attempted and how apt was it?).

Of particular importance in the Toulmin diagram (see Figure 2) is the "alternative explanations" category (which is located on the way up from the "data concerning student" or "data concerning situation" to the "claim about the student"). There are two main ways that our evaluations of performances may lead us astray. First, the situation may not include features that are needed in order to evoke the targeted knowledge. For example, if the goal of a training simulation is to assess students' proficiency to obtain information in an emotionally difficult patient history conversation, then selecting questions from a checklist and getting text responses misses a key feature for eliciting that skill. If the goal is merely knowing what questions to ask, that situation may be satisfactory; thereby reinforcing the point that one cannot know what features are critical without understanding the assessment's purpose and thus the kinds of claims it entails. Second, performing well on the task may require knowledge and skill beyond the targeted knowledge and skill. For example, lack of familiarity with a simulation interface can lead spuriously to poor performances. Messick (1989) refers to these two threats to validity as "construct underrepresentation" and "construct irrelevant demands." Messick's article (1994), "The Interplay of Evidence and Consequences in the Validation of Performance Assessments" remains the best source that explains how to think about what features should and should not be represented in a simulation. Every member of a team designing a simulation-based assessment should read Messick's piece—regardless of his or her area of expertise.

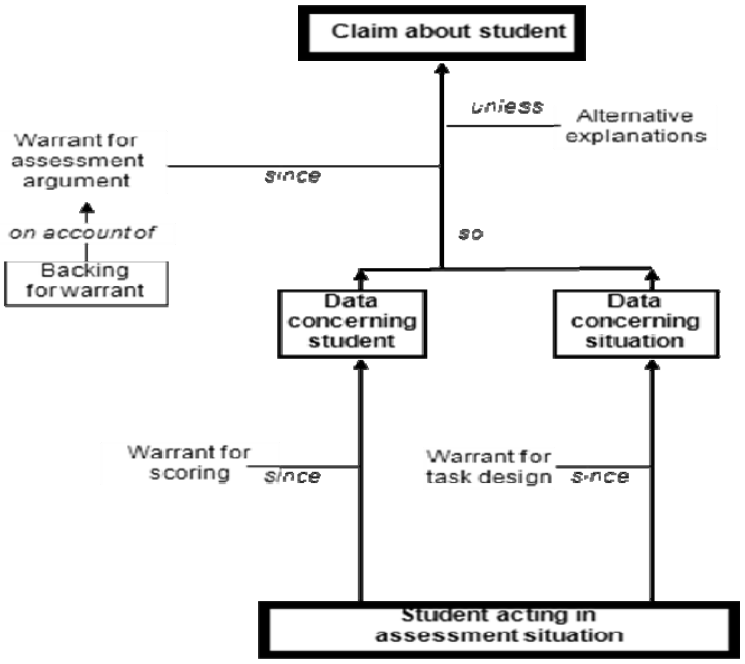

Figure 2. An Extended Toulmin argument diagram for assessment arguments

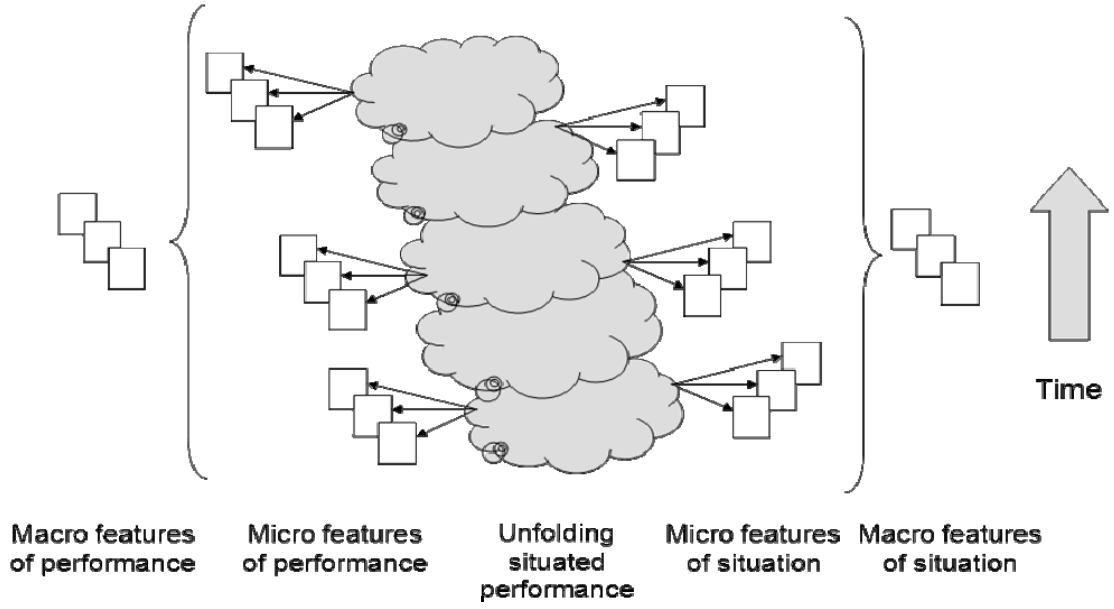

Figure 3. Potential points to identify data in performance assessments.

Working out assessment arguments for complex knowledge and skills is challenging; however, one of the tenets of ECD is that designers should not have to start every project anew. Assessment design patterns (Mislevy, et al., 2003) are a support tool adapted from architecture and engineering (e.g., Gamma et al., 1994), which are meant to sketch out a design space for certain hard-to-assess aspects of knowledge and skill in a way that (1) explicitly build around the elements of an assessment argument, (2) incorporate experience from research and previous experience, and (3) are flexible enough to help developers with different forms of assessment for a range of purposes. As an example, Table 1 is an abridged design pattern that supports the design of troubleshooting tasks.[2] It can guide creating specific tasks (multiple choice, written response, simulation, or actual-equipment tasks) or for determining what to seek evidence about in more complex tasks. Note that the results of the CTAs previously discussed can be abstracted to form the basis of a design pattern. For instance, in any domain in which design under constraints is relevant, the challenge of a task can be controlled in part by varying the number, explicitness, and degree of conflict of the constraints that must be addressed.

[2] See Mislevy, Riconscente, and Rutstein (2009) for a more detailed presentation of a suite of design patterns to support task design for model-based reasoning in scientific inquiry, and interactive online versions of the design patterns at
http://design-drk.padi.sri.com/padi/do/NodeAction?state=listNodes&NODETYPE=PARADIGMTYPE

Table 1.

A Design Pattern to Support the Assessment of Troubleshooting

Attribute	Value(s)
Name	Troubleshooting in a finite physical system (Related: Troubleshooting in an open system; network troubleshooting)
Overview	Built on hypothetico-deductive approach, using Newell-Simon model; e.g., problem space, active path, strategies such as serial elimination and space-splitting. This design pattern concerns evoking or identifying direct evidence about aspects of these capabilities in a given context.
Central claims	Capabilities in a specified context/domain to iteratively troubleshoot finite systems: propose hypotheses for system behavior, propose tests, interpret results, update model of system, identify and remediate fault.
Additional knowledge that may be at issue	Knowledge of system components, their interrelationships, and functions; Familiarity with tools, tests, and knowledge representations; Self-regulatory skills in monitoring progress.
Characteristic features	Situation presents system operating in accordance with fault(s). There is a finite (possibly very large) space of system states (cf. medical diagnosis). There are procedures for testing and repairing.
Variable task features	Complexity of system / Complexity of problem. Scope: Full problem with interaction; problem segment with interaction; problem segment with no interaction (e.g., multiple-choice hypothesis generation, explanation, or choose/justify next step). Setting: Actual system, interactive simulation, non-interactive simulation, talk-aloud, static representations Type of fault: Single v. multiple; constant or intermittent. Kind / degree of support: Reference materials (e.g., circuit diagrams, repair manuals); Advise from colleagues, real or simulated. Collaborative work? (If so, also use design pattern for collaboration)
Potential performances and work products	Final state of system; identification of fault(s); trace & time stamps of actions; video of actions; talk-aloud protocol; explanations or selections of hypotheses, choice of tests, explanations of test results, effects on problem space; constructed or completed representations of system at key points.
Potential features of performance to evaluate	*Regarding the final product:* Successful identification of fault(s)? Successful remediation? Total cost / time / number of actions. *Regarding performance:* Efficiency of actions (e.g., space-splitting when possible or serial elimination, vs. redundant or irrelevant actions); systematic vs. haphazard sequences of action. Error recovery. *Metacognitive:* Quality of self monitoring; quality of explanations of hypotheses, interpretation, selected actions.
Selected references	Newell & Simon (1972): Foundational reference on human problem-solving. Jonassen & Hung (2006): Cognitive model of troubleshooting. Steinberg & Gitomer (1996): Example with aircraft hydraulics.

The Conceptual Assessment Framework (CAF)

In the Conceptual Assessment Framework (CAF) the domain information is combined with information about goals, constraints, and logistics to create a blueprint for an assessment. The CAF comprises models whose objects and specifications provide the blueprint for the operational aspects of work, including (a) the creation of tasks, evaluation procedures, and statistical models, (b) delivery and operation of the assessment, and (c) analysis of data coming back from the field. Implementing these objects and coordinating their interactions in terms of the four-process delivery system (as described in the following sections) brings the assessment to life. While domain modeling emphasized the interconnections among aspects of people's capabilities, situations, and behaviors, the CAF capitalizes on the separability of the objects that are used to instantiate an assessment. Because the models and their components can be rendered in digital form, they are amenable to assisted and automated methods of generating, manipulation, operation, and assembly (Mislevy et al., 2010).

Figure 3 is a high-level schematic of the three central models in the CAF and the objects they contain. They are linked to each other through student-model variables, observable variables, work products, and task model variables—which formalize the elements in the assessment argument of Figure 2 (see Mislevy, Steinberg, & Almond, 2003, and Almond, Steinberg, & Mislevy, 2002).

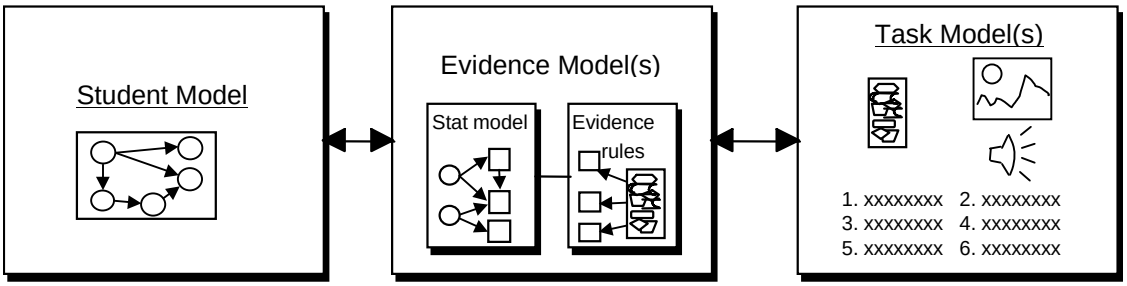

Figure 4. The Central Models of the Conceptual Assessment Framework.

The *Student Model* contains variables for expressing claims about targeted aspects of students' knowledge and skills. Their number and character depend on the purpose of the assessment (e.g., a single student-model variable to characterize students' overall proficiency in a domain of tasks for a certification decision; a multidimensional student model to sort out patterns of proficiency from complex performances or provide more detailed feedback). HYDRIVE used a multidimensional student model to track aspects of strategy knowledge,

familiarity with procedures, and knowledge of subsystems in the aircraft because its instructional decisions were made in these terms.

A *Task Model* formally describes the environment in which students say, do, or make something to produce evidence (see Vendlinski, Baker, & Niemi's 2008 study in which they provide templates and objects for authoring problem solving assessments). Regarding the actions and displays of the simulation environment to reactions to the examinee's actions, in simulation tasks more information is required than in fixed tasks. A common way of specifying a task is the initial status and transition rules of a finite state machine. A key design decision is specifying the form(s) in which students' performances will be captured, i.e., the Work Product(s)—for example, a sequence of steps in an investigation, the locations of icons dragged into a diagram, or the final solution of a design problem. The DISC CTA revealed that examinees' evaluations of diagnostic tests were more informative than their sequence of choices; thus, an additional work product consisting of filling out an insurance form with key findings and interpretations was recommended. Using underlying work-product data structures streamlines authoring, implementation, and evaluation (Luecht, 2009, Scalise & Gifford, 2006). Task model variables indicate salient features of a situation and are used in evaluating performances—data concerning the situation in the assessment argument. The values of some of the task model variables are set a priori by the task designer (e.g., What is the fault in the system? What is a patient's degree of gum recession?), while in interactive assessments dynamic task model variables are determined as the performance unfolds (e.g., Has the patient been stabilized? Given the examinee's troubleshooting actions so far, which components constitute the active path?)

An *Evidence Model* bridges the Student Model and the Task Model. The two components in an evidence model—evaluation and measurement—correspond to two steps of reasoning. The *evaluation component* delineates how one identifies and evaluates the salient aspects of the work products, which are expressed as values of Observable Variables. For instance, evaluation procedures can be rubrics for human scoring or algorithms for automated scoring procedures. Margolis and Clauser (2006) describe the regression-based automated scoring procedures used in the National Board of Medical Examiners' case management simulations; moreover, Braun, Bejar, and Williamson (2006) described the rule-based approach used for the Architectural Registration Examination site design problems. In simulations, evaluating Observable Variables may require information from dynamic task model variables. Efficiencies can again be gained through reuse and modular construction, as for example, different evaluation procedures can be used to extract different observable variables from the same work products when tasks are used for different purposes (or as

different ways of implementing procedures are used to extract the same observable variables from the same work products).

Data that are generated in the evaluation component are synthesized across tasks in the *measurement model* component. The simplest measurement models are classical test theory models, in which (possibly weighted) scores and subscores of salient features are added. Modular construction of measurement models assembles pieces of more complicated models such as those of item response theory or Bayesian inference networks (e.g., Mislevy & Levy, 2007). Of particular interest in simulations is assembling tasks and corresponding measurement models in accordance with task model variables (Martin & VanLehn, 1995; Mislevy & Gitomer, 1996; Shute, 2011). Much can be gained when the structure of evidentiary relationships in complex tasks and multivariate student models are expressed in re-usable measurement model fragments (Mislevy, Steinberg, Breyer, Johnson, & Almond, 2002). Using these schemas, task authors can create unique complex tasks but know ahead of time "how to score them."

It should be said that applying measurement modeling to simulation and game-based assessments is one of the frontiers of contemporary psychometrics (Rupp, Gushta, Mislevy, & Shaffer, 2010). While useful experience has been gained with a history of performance assessment, most of the practices and the language of measurement that are evolved for tests consist of discrete, pre-packaged tasks with just a few bits of data. Measurement researchers are extending the evidentiary reasoning principles that underlie familiar test theory to the new environment of the "digital ocean" of data (DiCerbo & Behrens, 2011).

Assessment Implementation

The Assessment Implementation layer of ECD is concerned with constructing and preparing the operational elements specified in the CAF. This includes authoring tasks, finalizing rubrics or automated scoring rules, estimating the parameters in measurement models, and producing simulation states and transition rules. Using common and compatible data structures increases opportunities for reusability and interoperability; furthermore, it helps bring down the costs of simulation-based assessment (see Chung, Baker, Delacruz, Bewley, Elmore, and Seely, 2008, on task design; Mislevy, Steinberg, Breyer, Almond, and Johnson, 2002, on measurement models; Luecht, 2002, on authoring frameworks; and Stevens and Casillas , 2006, on automated scoring).

Assessment Delivery

The Four-Process Architecture. The Assessment Delivery layer is where students interact with tasks, their performances are evaluated, and feedback and reports are produced.

Almond, Steinberg, and Mislevy (2002) lay out a four-process delivery system that can be used to describe computer-based testing procedures, as well as paper-and-pencil tests, informal classroom tests, and tutoring systems. When an assessment is operating, the processes pass messages among themselves in a pattern determined by the test's purpose. All of the messages are either data objects specified in the CAF (e.g., parameters, stimulus materials) or produced by the student (or other processes in data structures) that are specified in the CAF (e.g., work products, values of observable variables). Common language, data structures, and partitioning of activities promote the reuse of objects and processes as well as interoperability across projects and programs.

Figure 4 illustrates the four principal processes. The *activity selection process* selects a task or activity from the task library, or creates one in accordance with templates in light of what is known about the student or the situation. The *presentation process* is responsible for presenting the task to the student, managing the interaction, and capturing work products. Work Products are then passed on to the *evidence identification process*, or task-level scoring, which evaluates work using the methods specified in the Evidence Model. It sends values of Observable Variables to the *evidence accumulation process*, or test-level scoring, which uses the Measurement Models to summarize evidence about the student model variables and produce score reports. In adaptive tests this process provides information to the *activity selection process* to help determine what tasks to present next. A fixed-form multiple-choice test may require a single trip around the cycle. A simulation-based task can require many interactions among the processes in the course of a performance. For instance, Frezzo, Behrens, Mislevy, West, and DiCerbo (2009) describe the interplay among the four processes in the context of the simulation-based Packet Tracer Skills Assessment. Shute & Psotka, (1996) delineate how an intelligent tutoring system can jump out to instructional or practice models.

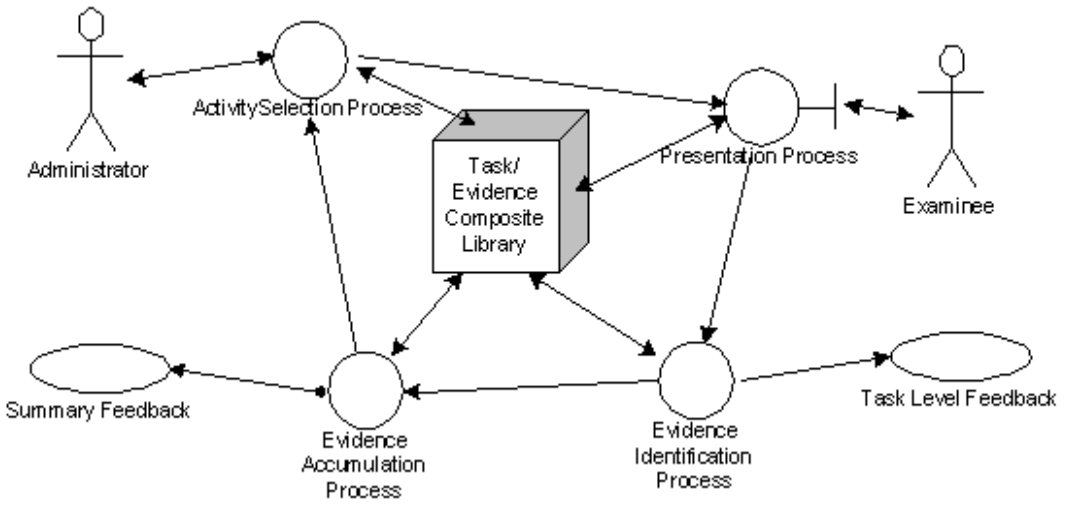

Figure 5. Processes in the assessment cycle.

Discussion

The discussion has touched upon designing simulations, simulation-based learning, and simulation-based assessments—all of these topics overlap substantially. In designing a simulation, attention is initially focused on the system that is being emulated. Whether an aspect of the human body, a mechanical system (e.g., a computer or airplane), or natural phenomena (e.g., an ecosystem), it is important to understand the simulation's elements, properties, behaviors, and responses to human (or other) actions. Although a simulation is inevitably less complete than the real-world system, in some directions it can be more developed (e.g., invisible phenomena made visible, stop-action and replay possible). The simulation designer faces many questions, such as what aspects to build into the simulation; what to simplify, modify, or ignore; and what affordances and interactions to provide users. These questions cannot be answered satisfactorily without identifying the purpose(s) of the effort.

When the focus of a simulation environment is learning, attention shifts to the cognitive and activity patterns that people develop in order to perform effectively in relevant situations. It is by developing and becoming skilled with these patterns that individuals overcome the pervasive difficulties that novices face in a broad range of domains (Salthouse, 1991) (e.g., knowing what is important, how to integrate information, the choice of action they may have, how to make/fluently carry out a choice of action). Other important questions are: In the project at hand, what aspects of knowledge and skill are the targets of learning? What skill level is the focus? What are the situations and possible actions that these users need in order

to develop their capabilities? How should practice opportunities and feedback be tuned to optimize their learning? Different answers to these questions can lead to very different choices for simulating aspects of the same domain. For example, The Microsoft Flight Simulator costs about $25 and helps beginners learn many of the essentials of flying quickly and enjoyably. A commercial airliner's full motion simulator costs millions of dollars, but it provides an authentic look and feel (down to the haptic properties of controls); when users are already highly trained professionals this is what is needed to advance their capabilities for extreme situations.

In other words, just because a simulation accurately emulates some aspects of a system does not necessarily mean that it will provide effective learning. A number of recent examples illustrate how designers have leveraged cognitive research to inform the design of simulations for learning. Frezzo (Frezzo, 2009; Frezzo, Behrens, & Mislevy, 2009), for instance, drew on activity theory (Engestrom, 1987) to design Packet Tracer interfaces to tune learning of various aspects of networking as a layer on top of code that simulates the behavior of networking devices. Working from a sociocognitive perspective, Shaffer (2006) built simulation environments that not only helped students learn knowledge and skills but also values, identities, and epistemologies that characterize professionals in domains (such as journalism and urban planning). It is important to move beyond emulating a system and concentrate on supporting students in learning how to interact with systems; yet, this requires serious thought and iterative cycles of design refinements and testing.

Moreover, moving beyond supporting learning to supporting assessment introduces an additional layer of design constraints. Just as learning science provides principles and methods for designing simulation environments for learning—assessment science provides principles and methods for designing simulation-based assessments. This report has provided an introduction to such an environment (namely that of evidence-centered assessment design); also, it has offered insights on using the framework for simulation-based assessment that draws on a number of projects.

A significant conclusion is that just because a simulation supports effective learning for some aspects of a system does not necessarily mean it will provide reliable or valid assessment of that learning. A great deal of thinking (for designing a simulation-based assessment in a certain domain) has gone into the design of an effective simulation for learning. In particular, the salient cognitive and activity patterns have been identified; features of situations that evoke this thinking and acting have been identified; and the kinds of things people do in these situations has been enabled. Doing this optimally to support learning is not equivalent to optimally providing evidence to an external observer, and being

able to characterize the properties of that evidence. For one thing, it is necessary to determine how to identify, evaluate, and synthesize relevant information in the examinee's actions. Everybody who wants to build a simulation-based assessment understands this. But not everybody understands that optimizing evidentiary value of simulation-based performances may require going back to the design of the simulator and the situations with fresh eyes—or at least through a psychometrician's perspective. Fewer options and more constrained situations, for example, may be less effective for learning but more effective for focusing examinees' actions on key aspects of cognitive or activity structures. Bringing floundering examinees back fairly early may not be as good for learning, but it is better for using limited assessment time. The requirement of an explicit work product may slow working through a simulation, but some valuable evidence regarding unobservable thinking will manifest. Simulation elements that provide authenticity for learning (e.g., working in the face of distractions) may introduce demands for knowledge and skill that are irrelevant to the purpose, and reduce its validity. Simply having lots of data, in the form of gigabytes of time-stamped mouse clicks and key strokes, may not provide much evidence if it is not about informative actions in relevant situations.

Moreover, a rich simulation setting that makes it possible to obtain information about many aspects of proficiency at the same time also makes it impossible to get very much reliable information about any of them. It can be effective to use different forms of assessment for different aspects of proficiency. The National Board of Medical Examiners, for example, uses multiple choice questions to assess a broad range of medical and scientific knowledge, coupled with computer simulations problems to get evidence about patient management and decision making over time, and live simulated patients to assess candidates' skills in person-to-person encounters. Tasks of each type stress the kinds of knowledge and skills they target and are less demanding than the ones that are addressed in the other tests.

This report has focused on the evidentiary-argument considerations that go into design decisions rather than psychometric methods for modeling the data. The author chose instead to highlight assessment-related issues that every member of the team can understand and should be responsive to. To understand exactly how different choices affect the evidentiary value of data, decision accuracy or instructional effectiveness will require more esoteric methods from the psychometrician's toolbox. The discussion of the Evidence Model in the CAF touched on these issues briefly and also cited some recent developments. More advances can be expected on this front but a long-standing lesson will remain applicable: To design a rich simulation environment, to collect data without consideration of how the data will be evaluated, and hoping psychometricians will somehow 'figure out how to score it' is

a bad way to build assessments. Close collaboration and interaction from the beginning of the design process is needed among users (who understand the purposes for which the assessment is intended), domain experts (who know about the nature of the knowledge and skills, the situations in which they are used, and what examinees do that provides evidence), psychometricians, (who know about the range of situations in which they can model data and examine and compare its evidentiary value), and software designers (who build the infrastructure to bring the assessment to life).

References

Almond, R. G., Steinberg, L. S., & Mislevy, R. J. (2002). Enhancing the design and delivery of assessment systems: A four-process architecture. *Journal of Technology, Learning, and Assessment, 1*(5). Retrieved from http://www.bc.edu/research/intasc/jtla/journal/v1n5.shtml.

Behrens, J.T., Mislevy, R.J., DiCerbo, K.E., & Levy, R. (in press). An evidence centered design for learning and assessment in the digital world. In M.C. Mayrath, J. Clarke-Midura, & D. Robinson (Eds.), *Technology-based assessments for 21st century skills: Theoretical and practical implications from modern research*. Charlotte, NC: Information Age.

Bejar, I.I., & Braun, H. (1999). Architectural simulations: From research to implementation: Final report to the National Council of Architectural Registration Boards (*ETS RM-99-2*). Princeton, NJ: Educational Testing Service.

Braun, H., & Bejar, I. I., & Williamson, D. M. (2006). Rule-based methods for automatic scoring: Application in a licensing context. In D. M. Williamson, R. J. Mislevy & I. I. Bejar (Eds.), *Automated scoring for complex constructed response tasks in computer based testing*. Mahwah, NJ: Lawrence Erlbaum Associates.

Cameron, C.A., Beemsterboer, P.L., Johnson, L.A., Mislevy, R.J., Steinberg, L.S., & Breyer, F.J. (1999). A cognitive task analysis for dental hygiene. *Journal of Dental Education, 64*, 333-351.

Chandler, P. & Sweller, J. (1991). Cognitive load theory and the format of instruction. *Cognition and Instruction, 8*, 293–332.

Chi, M. T. H., Glaser, R., & Farr, M. J. (eds.) (1988). The nature of expertise. Hillsdale, NJ: Erlbaum.

Chung, G.K.W.K., Baker, E.L., Delacruz, G.C., Bewley, W.L., Elmore, J., & Seely, B. (2008). A computational approach to authoring problem-solving assessments. In E.L. Baker, J. Dickieson, W. Wulfeck, and H.F. O'Neil (Eds.), *Assessment of problem solving using simulations* (pp. 289–307). Mahwah, NJ: Erlbaum.

Clarke-Midura, J., Code, J., Dede, C., Mayrath, M., & Zap, N. (in press). Thinking outside the bubble: Virtual performance assessments for measuring complex learning. In D. Robinson, J. Clarke-Midura, & M. Mayrath (Eds.), *Technology-based assessments for 21st century skills: Theoretical and practical implications from modern research.* Charlotte, NC: Information Age.

DiCerbo, K. E., & Behrens, J. T. (2011). Implications of the digital ocean on current and future assessment. In R. Lissitz & H. Jiao (Eds.), *Computers and their impact on state assessment: Recent history and predictions for the future*. Charlotte, NC: Information Age Press.

Dillon, G.F., Boulet, J.R., Hawkins, R.E., & Swanson, D.B. (2004). Simulations in the United States Medical Licensing Examination™ (USMLE™). *Quality and Safety in Health Care, 13* (Supplement 1), 41-45.

Engeström, Y. (1987). *Learning by expanding: An activity theoretical approach to developmental research*. Helsinki, Finland: Orienta Konsultit.

Ericsson, A.K., Charness, N., Feltovich, P., & Hoffman, R.R. (2006). *Cambridge handbook on expertise and expert performance*. Cambridge, UK: Cambridge University Press.

Ewing, M., Packman, S., Hamen, C., & Thurber, A.C. (2010) Representing targets of measurement within evidence-centered design. *Applied Measurement in Education, 23,* 325-341.

Flanagan, J. C. (1954). The critical incident technique. *Psychological Bulletin, 51,* 327-359.

Frezzo, D. C. (2009). *Using activity theory to understand the role of a simulation-based interactive learning environment in a computer networking course*. Unpublished doctoral dissertation, University of Hawai'i, Honolulu, Hawai'i.

Frezzo, D. C., J. T. Behrens, & R. J. Mislevy. (2009). Activity theory and assessment theory in the design and understanding of the packet tracer ecosystem. *The International Journal of Learning and Media, 2*. Retrieved from http://ijlm.net/knowinganddoing/10.1162/ijlm.2009.0015

Frezzo, D.C., Behrens, J.T, Mislevy, R.J., West, P., & DiCerbo, K.E. (2009). Psychometric and evidentiary approaches to simulation assessment in Packet Tracer software. *ICNS '09: Proceedings of the fifth international conference on networking and services* (pp. 555-560). Washington, DC: IEEE Computer Society.

Gamma, E., Helm, R., Johnson, R., & Vlissides, J. (1994). *Design patterns: Elements of reusable object-oriented software.* Reading, MA: Addison-Wesley.

Greeno, J. G. (1998). The situativity of knowing, learning, and research. *American Psychologist, 53,* 5–26.

Haertel, G., Wentland, E., Yarnall, L., & Mislevy, R.J. (in press). Evidence-centered design in assessment development. In C. Secolksy & B. Denison (Eds.), *Handbook of measurement, assessment, and evaluation in higher education.* New York, NY: Routledge.

Huff, K., Steinberg, L., & Matts, T. (2010). The promises and challenges of implementing evidence-centered design in large-scale assessment. *Applied Measurement in Education. 23,* 310-324.

Jonassen, D.H., & Hung, W. (2006). Learning to troubleshoot: A new theory-based design architecture. *Educational Psychology Review, 18,* 77-114

Katz, I.R. 1994. Coping with the complexity of design: Avoiding conflicts and prioritizing constraints. In A. Ram, N. Nersessian, and M. Recker (Eds.), *Proceedings of the sixteenth annual meeting of the Cognitive Science Society*, (pp.485-489). Mahwah, NJ: Erlbaum.

Luecht, R. M. (2002). *From design to delivery: Engineering the mass production of complex performance assessments*. Paper presented at the annual National Council on Measurement in Education meeting, New Orleans, LA.

Luecht, R.M. (2009). Adaptive computer-based tasks under an assessment engineering paradigm. In D. J. Weiss (Ed.), *Proceedings of the 2009 GMAC conference on computerized adaptive testing.* Retrieved from www.psych.umn.edu/psylabs/CATCentral/

Margolis, M. J., & Clauser, B. E. (2006). A regression-based procedure for automated scoring of a complex medical performance assessment. In D. Williamson, R. Mislevy & I. Bejar (Eds.) *Automated scoring of complex tasks in computer based testing* (pp. 123-167). Hillsdale, NJ: Lawrence Erlbaum Associates.

Martin, J.D., & VanLehn, K. (1995). A Bayesian approach to cognitive assessment. In P. Nichols, S. Chipman, & R. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 141-165). Hillsdale, NJ: Erlbaum.

Melnick, D. (1996). The experience of the National Board of Medical Examiners. In E.L. Mancall, P.G. Vashook, & J.L. Dockery (Eds.), *Computer-based examinations for board certification* (pp. 111-120). Evanston, IL: American Board of Medical Specialties.

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

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (*3rd ed.*, pp. 13-103). New York, NY: American Council on Education/Macmillan.

Mislevy, R.J. (2003). Substance and structure in assessment arguments. *Law, Probability, and Risk, 2,* 237-258.

Mislevy, R.J., Behrens, J.T., Bennett, R.E., Demark, S.F., Frezzo, D.C., Levy, R., Robinson, D.H., Rutstein, D.W., Shute, V.J., Stanley, K., & Winters, F.I. (2010). On the roles of external knowledge representations in assessment design. *Journal of Technology, Learning, and Assessment, 8*(2). Retrieved from http://escholarship.bc.edu/jtla/vol8/2.

Mislevy, R.J., & Gitomer, D.H. (1996). The role of probability-based inference in an intelligent tutoring system. *User-Modeling and User-Adapted Interaction, 5,* 253-282.

Mislevy, R.J., Hamel, L., Fried, R.G., Gaffney, T., Haertel, G., Hafter, A., Murphy, R., Quellmalz, E., Rosenquist, A., Schank, P., Draney, K., Kennedy, C., Long, K., Wilson, M., Chudowsky, N., Morrison, A., Pena, P., Songer, N. and Wenk, A. (2003). *Design patterns for assessing science inquiry (PADI Technical Report 1),* Menlo Park, CA: SRI International. Retrieved from http://padi.sri.com/downloads/TR1_Design_Patterns.pdf.

Mislevy, R. J., & Levy, R. (2007). Bayesian psychometric modeling from an evidence-centered design perspective. In C. R. Rao and S. Sinharay (Eds.), *Handbook of statistics, 26,* (pp.839-865). North-Holland: Elsevier.

Mislevy, R.J., & Riconscente, M.M. (2006). Evidence-centered assessment design: Layers, concepts, and terminology. In S. Downing & T. Haladyna (Eds.), *Handbook of Test Development* (pp. 61-90). Mahwah, NJ: Erlbaum

Mislevy, R. J., Riconscente, M. M., & Rutstein, D. W. (2009). *Design patterns for assessing model-based reasoning.* (Large-scale Assessment Technical Report 6). Menlo Park, CA: SRI International. Retrieved from http://ecd.sri.com/downloads/ECD_TR6_Model-Based_Reasoning.pdf.

Mislevy, R.J., Steinberg, L.S., & Almond, R.A. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspective*s, 1, 3-67.

Mislevy, R.J., Steinberg, L.S., Breyer, F.J., Johnson. L., & Almond, R.A. (2002). Making sense of data from complex assessments. *Applied Measurement in Education, 15,* 363-378.

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

Pearlman, M. (2004). The design architecture of NBPTS certification assessments. In R.E. Stake, S. Kushner, L. Ingvarson, & J. Hattie (Eds.) *Assessing teachers for professional certification: The first decade of the National Board for Professional Teaching Standards (advances in program evaluation), 11,* 55–91. United Kingdom: Emerald Group Publishing Limited.

Roschelle, J. (1997). Designing for cognitive communication: Epistemic fidelity or mediating collaborative inquiry? In. D. L. Day & D. K. Kovacs (Eds.) *Computers, communication, and mental models.* Bristol, PA: Taylor & Francis.

Rupp, A.A., Gushta, M., Mislevy, R.J., & Shaffer, D.W. (2010). Evidence-centered design of epistemic games: Measurement principles for complex learning environments. *Journal of Technology, Learning, and Assessment, 8*(4). Retrieved from http://www.jtla.org.

Salthouse, T.A. (1991). Expertise as the circumvention of human processing limitations. In K.A. Ericcson & J. Smith (Eds.), *Toward a general theory of expertise*, (pp. 286-300). Cambridge, England: Cambridge University Press.

Scalise, K., & Gifford, B. (2006). Computer-based assessment in e-learning: A framework for constructing "intermediate constraint" questions and tasks for technology platforms. *Journal of Technology, Learning, and Assessment, 4*(6). Retrieved from http://www.jtla.org.

Schraagen, J.M., Chipman, S.F., & Shalin, V.J. (2000). *Cognitive task analysis.* Mahwah, NJ: Erlbaum.

Shaffer, D.W. (2006). *How computer games help children learn.* New York: Palgrave/Macmillan.

Shute, V. J. (2011). Stealth assessment in computer-based games to support learning. In S. Tobias & J. D. Fletcher (Eds.), *Computer games and instruction* (pp. 503-524). Charlotte, NC: Information Age Publishers.

Shute, V. J., & Psotka, J. (1996). Intelligent tutoring systems: Past, present, and future. In D. Jonassen (Ed.), *Handbook of research for educational communications and technology* (pp. 570-600). New York, NY: Macmillan.

Snir, J., Smith, C., & Grosslight, L. (1993). Conceptually enhanced simulations: A computer tool for science teaching. *Journal of Science Education and Technology, 11,* 373-388.

Steinberg, L. S., & Gitomer, D. H. (1996). Intelligent tutoring and assessment built on an understanding of a technical problem-solving task. *Instructional Science, 24*, 223-258.

Stevens, R., & Casillas, A. (2006). Artificial neural networks. In D.M. Williamson, R.J. Mislevy, & I.I. Bejar (Eds.), *Automated scoring of complex tasks in computer based testing* (pp. 259-312). Mahwah, NJ: Erlbaum Associates.

Toulmin, S.E. (1958). *The uses of argument.* Cambridge: Cambridge University Press.

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

Vygotsky, L.S. (1978). *Mind and society: The development of higher psychological processes.* Cambridge, MA: Harvard University Press.

Wertsch, J. (1998). *Mind as action*. New York: Oxford University Press.

Williamson, D.M., Bauer, M., Steinberg, L.S., Mislevy, R.J., Behrens, J.T., & DeMark, S. (2004). Design rationale for a complex performance assessment. *International Journal of Measurement, 4,* 303-332.