CRESST REPORT 778

John T. Behrens
Robert J. Mislevy
Kristen E. DiCerbo
Roy Levy

AN EVIDENCE CENTERED DESIGN FOR LEARNING AND ASSESSMENT IN THE DIGITAL WORLD

DECEMBER, 2010

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

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

An Evidence Centered Design for Learning and Assessment in the Digital World

CRESST Report 778

John T. Behrens
Cisco

Robert J. Mislevy
University of Maryland

Kristen E. DiCerbo
Independent Researcher

Roy Levy
Arizona State University

December, 2010

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 Educational Research and Development Center Program, PR/Award Number R305C080015, as administered by the Institute of Education Sciences, U.S. Department of Education.

The findings and opinions expressed here do not necessarily reflect the positions or policies of the National Center for Education Research (NCER), the Institute of Education Sciences (IES), or the U.S. Department of Education.

This report is adapted with permission from the authors' chapter in Mayrath, M.C., Clarke-Midura, J., & Robinson, D. (Eds.) (2011), *Technology-based assessments for 21st century skills: Theoretical and practical implications from modern research*. Charlotte, NC: Information Age.

To cite this report, please use the following as your APA reference: Behrens, J.T., Mislevy, R.J., DiCerbo, K.E., & Levy, R. (2010). *An evidence centered design for learning and assessment in the digital world.* (CRESST Report 778). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

TABLE OF CONTENTS

EVIDENCE CENTERED DESIGN FOR LEARNING AND ASSESSMENT IN THE DIGITAL WORLD

John T. Behrens
Cisco

Robert J. Mislevy
University of Maryland

Kristen E. DiCerbo
Independent Researcher

Roy Levy
Arizona State University

Abstract

The world in which learning and assessment must take place is rapidly changing. The digital revolution has created a vast space of interconnected information, communication, and interaction. Functioning effectively in this environment requires so-called 21st century skills such as technological fluency, complex problem solving, and the ability to work effectively with others. Unfortunately, traditional assessment models and methods are inadequate for evaluating or guiding learning in our digital world. This report argues that the framework of evidence-centered assessment design (ECD) supports the design and implementation of assessments that are up to the challenge. We outline the essential ECD structure and discuss how the digital world impacts each phase of assessment design and delivery. The ideas presented in the report are illustrated with examples from our ongoing experiences with the Cisco Networking Academy. We have used this approach to guide our work for more than 10 years and ultimately seek to fundamentally change the way networking skills are taught and assessed throughout the world, including the delivery of 100 million exams in over 160 countries and innovative simulation-based curricular and assessment tools.

Introduction

If the 21st century unfolds similarly to previous centuries, we can be certain that time will be uniformly distributed; technological and social transformations will increase exponentially; and almost any attempt at predicting further change will underestimate the amount of actual change that will occur over the next 90 years. What appears most salient about the 21st century, in its current nascent state, are the individual as well as societal changes brought about by the Information and Communication Technologies (ICT) of digitization, computation, and information transmission via communications networks such as the World Wide Web (WWW). Consider, for instance, how you would have completed undertaking each of the following activities in 1990:

- 1. Purchase a shirt from a company 1,000 kilometers away for whom you do not know the address or phone number.
- 2. Determine the height of the St. Joseph River in Elkhart, Indiana today.
- 3. Show someone living on another continent, in real time, what your child looks like when dancing.

At the present time, these tasks would be considered relatively simple because of the ubiquity of digitization devices (cameras, remote sensors); computation of digital information; the transmission of information via computer as well as other information networks; and display via the World Wide Web (WWW). Twenty years ago, each of these tasks would have been difficult to complete. They would probably require expensive and time-intensive physical movement or access to information previously held by proprietary groups (the local phone company). Currently because of technological advances, the information could be acquired in the public domain. Search engines (e.g., Google.com; Ask.com) now provide global contact information. NOAA.gov, for instance, provides data sensors to thousands of rivers and creeks in the United States and numerous free internet-based video chatting services are also available. Technologies interacting through the WWW allow us to see into homes and schools around the world, visualize data from space, and talk face to face with a colleague in another country.

We will refer to this breadth of technological advances as the digital revolution (DR). At the present time, technologies advance so rapidly and have become so commonplace that we hardly notice. These advances change the ways in which we are able to assess knowledge, skills, and attributes (KSAs); what we perceive as relevant to assess; and how we think about the very nature of assessment. In this report, we will discuss how we might understand the impacts of technological and social shifts in terms of the Evidence Centered Design (ECD) (Mislevy, Steinberg, & Almond, 2003) conceptual framework for assessment. We have been using this approach in our work for over 10 years to undergird the delivery of 100 million exams in over 160 countries, along with development of innovative simulation-based curricular and assessment tools (e.g., Frezzo, Behrens & Mislevy, 2010). The scope and scale of such work would have probably not been imagined 20 years ago. For each of the major sections of the ECD framework, we offer thoughts about how emerging technologies will influence the future of assessment and provide examples from our own emerging work.

History and Context

The context of our assessment research and experience is the Cisco Networking Academies (CNA; see http://cisco.com/go/netacad/), a public-private partnership between Cisco and over 9,000 educational institutions in over 160 countries. Cisco, previously called Cisco Systems, is the world's largest manufacturer of computer and data networking hardware and related equipment. Cisco provides partnering schools with free online curriculum and online assessments to support local school instructors in teaching ICT skills in areas related to PC repair and maintenance, as well as computer and data network design, configuration, and maintenance in alignment with entry-level industry certifications. The value of the program from the perspective of corporate social responsibility was discussed by Porter and Kramer (2002) while the logical origins of the e-learning approach have been described by Levy and Murnane (2005). Behrens, Collison, and DeMark (2005) provide a conceptual framework for the many and varied aspects of the assessment ecosystem in the program.

The instructional cycle in the Networking Academies typically consists of the students working through the interactive online curriculum prior to class time. This is followed by classroom face-to-face interaction, which provides the opportunity for group activities; additional clarification by the instructor; and hands-on experience with networking equipment. The e-learning environment includes facilities for simulations of networks that would prove too complex or varied to operate with hardware in the classroom (Frezzo et al. 2010). In order to be successful in this domain, students must learn a broad range of planning, design, implementation, operating, and troubleshooting skills that combine a conceptual understanding of how networking systems work. Students must also familiarize themselves with the physical aspects of network connectivity (such as care and organization of cables, alignment of hardware in physical spaces) and facility with the programming language of computer and data networks called the Cisco IOS (Frezzo, Behrens, & Mislevy, 2009). Student-initiated formative assessment and curriculum-embedded feedback occur throughout the learning progression with built-in interactive curricular objects; in-line fixed-response quizzes; simulation-based challenge labs, which simulate complex tasks and provide performance feedback; and numerous simulation-based practice activities. In addition, a separate online assessment system provides instructor initiated assessments for the end-of each chapter; end-of course fixed-response exams; and end-of-course simulation-based performance exams.

Why ECD?

In 2000, the Networking Academies undertook a two-pronged effort to advance its nascent assessment program. On the one hand, there were efforts to redesign the large-scale fixed-response assessment system. This system was used for students' chapter and final exams as they progressed through the learning experience. On the other hand, a new strand of work was initiated to investigate the possibility of automated performance-based assessment. This work eventually produced the NetPass system (Behrens, Mislevy, Bauer, Williamson, & Levy, 2004; Williamson et al., 2004). At first, the primary concern was to balance the need for a framework that could be implemented in a fairly standard way (e.g., a large-scale multiple-choice testing program). The framework also had to be abstract enough so that it could be extended as needed in the very unclear future. In the second phase, the primary concern was that traditional assessment language was inadequate for the job. Assessment designers and teachers involved in the construction of the program could easily create real-world tasks and corresponding scoring systems with ostensibly high ecological validity. However, the open-ended structure of the work was a poor fit for the fixed-response oriented language and technologies that ground familiar large-scale testing programs. Presenting a learner with a computer network and asking her to fix it is a relevant and straightforward task. Yet we were at a loss when deciphering how to match these aforementioned tasks, which occur naturally in the professional world, with the need for automated scoring in the global online educational system. Where does the language of question and answer or of a correct and incorrect response fit into the fluid, seamless, and interactive experience of working on computer networking equipment? There is no question—only a task and there is no answer— only a working (or not working) network. Options and distracters were likewise difficult to map onto this environment. In short, we required a language that subsumed fixed-response tasks but did not constrain us to them.

Relevant experience and research was available from studies of performance assessment in the field of Education (e.g., Kane & Mitchell, 1996); simulation testing in professional settings (e.g., Tekian, McGuire, & McGahie, 1999); and intelligent tutoring systems with implicit assessments of students' capabilities (e.g., Shute & Psotka, 1996). Furthermore, theoretical pieces such as Messick (1994) as well as Wiley and Haertel (1996) began to present a way of thinking about assessment that could unify the principles that underlie assessment of all forms-- from multiple-choice and performance tasks to extended projects and informal interactions with students. A quotation from Messick (1994) neatly summarizes the core idea of an assessment argument:

> A construct-centered approach 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. 17).

To the end of instantiating such an argument, the ECD framework provides terminology and representations for layers at which fundamental entities, relationships, and activities continually appear in assessments of all kinds. We posit that the essential structure will remain useful for thinking about assessments even as technology and psychology advance. However, it becomes apparent that these advances can provoke radically different answers to Messick’s questions about the nature of what to assess; what situations hold evidence; which performances should be observed; and how to evaluate them. ECD, nevertheless, provides a common conceptual framework for designing and deciphering assessments across a range of surface forms.

In the case of Cisco Networking Academies, ECD has served us by providing a sufficiently high degree of abstraction in the language to encompass standard practice. While presenting a set of constructs that are equally well applied to describe teacher-student interactions, ECD also provides a wide array of interactions with complex automated assessment systems. A key feature attributed to ECD’s utility to our work has been that the model is fundamentally descriptive rather than prescriptive. Where the ECD model states that an evidence model bridges the statistical model and the task model, it is not asserting that you should construct an evidence model of any particular kind; rather, it highlights that the ideal way to move from a task to an inference is to use an evidence model to reason from observation to conclusion-- no matter how informal or implicit the activity. This step may be explicit and technical in one assessment and implicit and intuitive in another but recognizing it as an essential link in a web of reasoning of all assessments helps us understand existing assessments and assists in the design of new ones. We recognize our actions when implicit, and know what kinds of structures and relationships will need to be in place when we need to make them explicit. Such conceptualizations provide a comprehensive language for some of the unarticulated steps in common assessment development work. In instances where test developers might have said, "It works" or "Just make some items and see what sticks," ECD provides a more detailed and explicit language to help us understand the ways our activities are working and for which purposes. As technologies and the social changes around them accelerate at an increasing pace, we have found that the value of ECD increases as people attempt to make sense of technologies and their possible uses.

The ECD Framework and its relationship to technology

Educational assessment involves characterizing aspects of student knowledge, skill, or other attributes (KSAs), based on inferences from the observation of what they say, do, or make in certain kinds of situations. The following sections describe key parts of the ECD framework as end-to-end processes in several conceptual layers (Mislevy, 1994; Mislevy, Steinberg, & Almond, 2002, 2003). The first step in starting the assessment process is considering characteristics of the world that are relevant to the assessment one wishes to construct. This is represented by the top layers of the ECD model, as illustrated in Figure 1 (adapted from Mislevy & Riconscente, 2006): The first layer is marshalling facts and theory about the domain and the second is organizing the information in the form of assessment arguments. The middle layer, the Conceptual Assessment Framework (CAF), specifies more technical models for task creation; evaluation procedures; measurement models; and the like—in essence, blueprints for the pieces and activities that instantiate the argument in the real world. The next layer concerns the manufacturing of the assessment artifacts and the specifics for their usage. The lower layer describes a four-process architecture for understanding assessment delivery.

Below we discuss the major elements of the ECD framework and note how technological advances affect the conceptualization of the element and/or its implementation in an assessment system. In some areas the technological directions are clear; in other areas, they are more speculative.

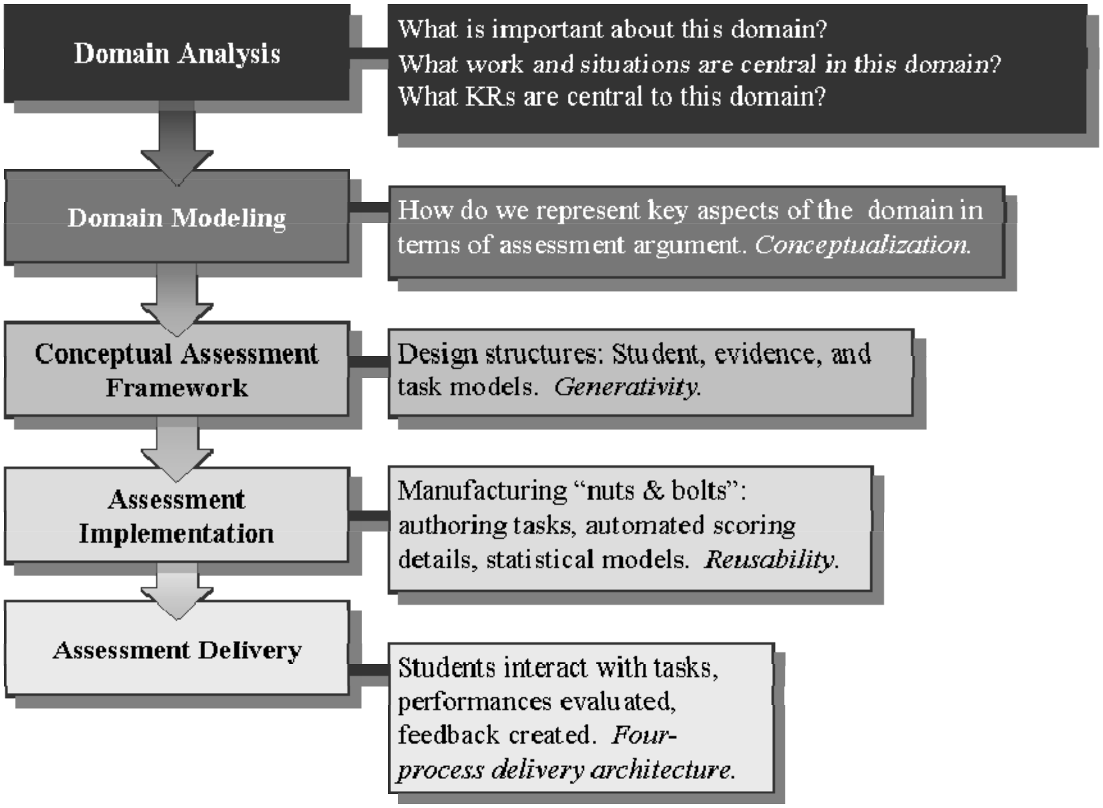

Domain Analysis
What is important about this domain?
What work and situations are central in this domain?
What KRs are central to this domain?
Domain Modeling
How do we represent key aspects of the domain in terms of assessment argument. Conceptualization.
Conceptual Assessment Framework
Design structures: Student, evidence, and task models. Generativity.
Assessment Implementation
Manufacturing "nuts & bolts": authoring tasks, automated scoring details, statistical models. Reusability.
Assessment Delivery
Students interact with tasks, performances evaluated, feedback created. Four-process delivery architecture.

Figure 1. Layers in Evidence-Centered Design (adapted from Mislevy & Riconscente, 2006).

Domain Analysis

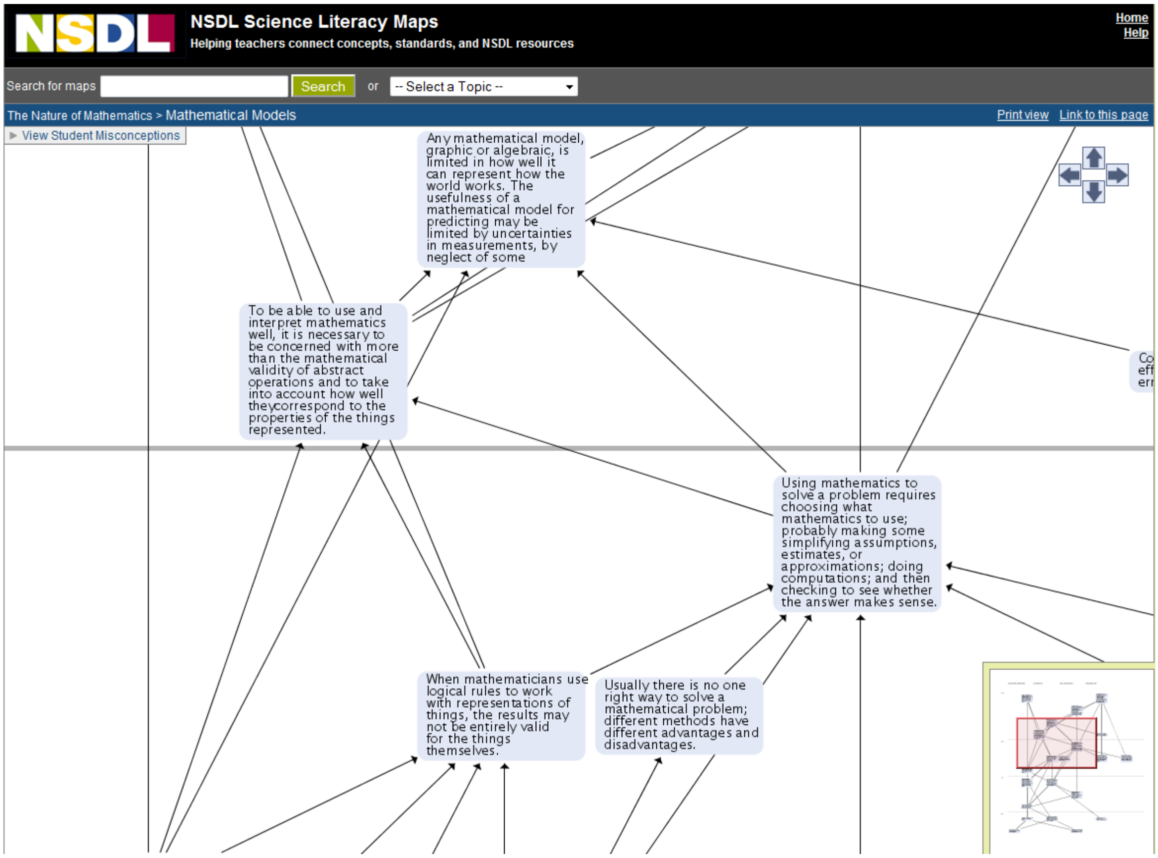
Domain analysis and domain modeling define the high-level content or experiential domains to be assessed and documents relationships among them. This is the content knowledge or subject matter related to the educational enterprise; the ways people use it; and the kinds of situations they use it in. What 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? For the current context, what knowledge, skills, goals, tools, enablers, representations and possible distractions are relevant? How do people interact with the physical environment, conceptual representations, and other individuals to accomplish certain feats? By what standards are these efforts judged? To answer such questions, designers conduct a domain analysis. They 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; test purposes; and the various requirements, resources and constraints to which the proposed assessment might be subject.

Developments in psychology have had profound effects on domain analysis in recent decades (Mislevy, 2006). For example, domain analysis under behaviorist psychology, focused on identifying concrete and precisely-defined actions in a domain, to be expressed as behavioral objectives (Mager, 1962). The information-processing perspective of the cognitive revolution (Newell & Simon, 1972) called attention to the internal and external knowledge representations that people work with; the procedures and strategies they use; and the features of problems that make them hard--all holding clear implications for assessment design. A sociocognitive perspective further widens domain analysis to the range of cognitive, cultural, and physical tools people use and the ways they interact with situations and each other to accomplish goals in some sphere of activity (e.g., Engeström, 1999). Domain analyses carried out under the latter two perspectives increasingly reveal the importance of peoples interaction with the technological environment noted in the introduction. Thus, technology continually gives rise to new forms in the representation and communication of extant information in terms of knowledge representations (KRs). Moreover, technology creates new and oftentimes more complex tools and environments that people must attune themselves to, in order to create and transform information. In turn, this gives rise to the kinds of capabilities we need to assess. The ability to designing and troubleshoot computer networks is a prototypical example of an important domain that did not exist until recently.

In times of rapid economic, social, and political change, the existence and composition of domains will change rapidly. For instance, prior to the digital revolution, office secretaries in a typing pool and stenographers were a common vocational track. These jobs have largely disappeared with the advent of personal computers and the ubiquity of typing skills in mature economies. The aforementioned changes have socio-political consequences because the educational and professional constituents often vary in the abstractness or generalizability in the skill and knowledge of focus. Professional training, staffing, and guild organizations often support workers in acquiring needed abilities when there is a rapid change in the demand for certain work force skills. Educational organizations generally focus on broader educational shifts. Accordingly, periods of rapid societal change can increase the divergence of short and long term foci, which can thereby increase friction in educational and assessment rhetoric. For example, in our contemporary and rapidly changing environment, there is a contrast between the conceptualization of skill and knowledge (as they would impact workforce retraining) against the broader capabilities that general education is meant to develop. As a result, assessment designers need conceptual models that can accommodate these types of variation in domain focus and proficiency models.

While domain analysis and its monitoring has traditionally been done via human analysis of job tasks, job requirements, and similar artifacts, there has been significant growth in the application of computer-based semantic analysis (Baayen, 2008; Biber, Conrad, & Reppen, 1998; Gries, 2009; Manning & Schuetze, 1999) to aid in the extraction of patterns in electronic data that may suggest shifts in domains or emergence of new domains. Such semantic analysis has historically been conducted by humans reading text but as the artifacts of human activity (work products) become increasingly digital, the door is open for automated techniques to identify trends in new activity. In 2008, a white paper published by International Data Corporation (IDC) entitled "The Diverse and Exploding Digital Universe," argued that the digital universe (the size of all electronic data) was 10 percent larger than previously estimated, and that by 2011 it will be 10 times larger than it was in 2006 (IDC, 2008). A report by the National Academies notes that particle physics experiments conducted with the Large Hadron Collider at CERN are expected to generate 15 petabytes of data annually, thereby matching the amount of information stored in all U.S. academic and scientific libraries every two years (National Academy of Science, 2009). These technological shifts in data collection will drive the need for new skills and methods for approaching science in the computer age (e.g., Wolfram, 2002); compatible new methods for revising; and tracking changes to the structure and content of domains. As the core representations and understandings of domains quickly evolve, the challenges and opportunities for moving those understandings and representations into the educational and assessment world will continue to increase as well.

One method for addressing the need for unified understandings of domains is the application of WWW technologies for the consolidation and distribution of domain models. Consider the digitization of the Atlas of Science Literacy (AAAS 2001; AAAS, 2007) provided by the National Science Digital Library (http://strandmaps.nsdl.org/). These online representations communicate a number of important attributes of specific science concepts by placing them in a graphical space of grade level (vertical placement) and conceptual strand (horizontal placement) while indicating pre-requisite or supporting relationships with arrows. The center panel is a detailed view of the subsection of the entire model which is depicted in the lower right panel. Each node in the model has hyperlinks to additional information and an overlay of relevant student misconceptions can seen by using the pull-down menu in the upper left corner. As organizations evolve to use such distributed displays, the assessment community will benefit by maintaining a united and re-useable set of representations from which to carry out domain modeling and subsequent assessment artifacts.


NSDL
NSDL Science Literacy Maps
Helping teachers connect concepts, standards, and NSDL resources
Home
Help
Search for maps
Search
or
-- Select a Topic --
The Nature of Mathematics > Mathematical Models
Print view
Link to this page
View Student Misconceptions
Co
eff
err

Any mathematical model, graphic or algebraic, is limited in how well it can represent how the world works. The usefulness of a mathematical model for predicting may be limited by uncertainties in measurements, by neglect of some

To be able to use and interpret mathematics well, it is necessary to be concerned with more than the mathematical validity of abstract operations and to take into account how well theycorrespond to the properties of the things represented.

Using mathematics to solve a problem requires choosing what mathematics to use; probably making some simplifying assumptions, estimates, or approximations; doing computations; and then checking to see whether the answer makes sense.

When mathematicians use logical rules to work with representations of things, the results may not be entirely valid for the things themselves.

Usually there is no one right way to solve a mathematical problem; different methods have different advantages and disadvantages.

Figure 2. Science Literacy Map of the Mathematical Models domain. Interactive graphics available at http://strandmaps.nsdl.org.

Domain Modeling

In domain modeling, designers organize information from domain analyses to describe relationships among capabilities, what we might see people say, do, or make as evidence, and situations and features to evoke it—in short, the elements of assessment arguments. Graphical and tabular representations and schemas are constructed to convey these relationships. Furthermore, prototypes may be used to fix ideas or test assumptions. Among the representational forms that have been used to implement ECD are claims and evidence worksheets; Toulmin diagrams for assessment arguments; 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). A sample of a claims and evidence form from a CNA curriculum regarding routing is shown in Figure 3. These can serve as targets for creating specific tasks (multiple choice, written response, simulation, or actual-equipment tasks) or for determining what to seek evidence about in more complex tasks that encompass the specified claim as well as potentially several other claims. The next section will give an example of a design pattern.

Claim 402: Develop a design that incorporates specified new device(s) into a network with minimum disruption to network operation

Representations to convey information to student (some or all to be presented to student):

Scenario (includes system requirements); building diagram; (existing) network diagram; existing configuration; physical network; simulation of network

Essential features (to be specified in task):

Timelines; device(s) to be added; characteristics of the network; location of utilities; telecommunications; distances; applications

Representations to capture information from student (e.g., potential work products):

Materials list; (final) network diagram; chart of IP addressing and subnet masking; cut sheet; number-base conversions worksheet; backup plan

Observable Features (i.e., evidence that can be identified from work products):

Re documentation: Completeness; accuracy

Re proposed solution: Practicality; cost effectiveness; timeliness; effective/appropriate use of existing assets; migration strategy

Re procedures: Efficiency; total time; down time

Figure 3. Example of a Claims and Evidence form.

21st Century Skills

A first challenge for assessing 21st century skills lies in the area of domain analysis and modeling. When people use the phrase, 21st century skills, just what capabilities are they referring to? What do we know about the development and the performance of those skills in real-world situations? The idea of 21st century skills is currently a rather amorphous theoretical construct suggesting new domains for occupational, educational, and personal activity. These domains are in need of more thorough analysis and modeling.

The term 21st century skills has come to be associated with more broadly defined notions of communication, collaboration, and problem-solving – all of which remain important as environments change but take forms shaped by the those environments. Fixing a Model-T is a substantially different cognitive activity from trouble-shooting a computer network. However, there are pervasive principles and structures that can be adduced to help design instruction and assessment and more importantly, ground students' learning so it can adapt to the situations of tomorrow (Schaafstal & Schraagen, 2000). Developing design patterns at this level can impart meaning to 21st century skills, in ways that can guide practical assessment development. A design pattern creates a design space to help task developers think through options and choices in designing tasks to evoke evidence of some targeted aspect of learners' capabilities or to recognize and evaluate the evidence as it arises in less structured assessments, such as simulations and games.

Drawing on Wise Rutstein's study (2005), Table 1 shows an abbreviated design pattern to support assessment design for troubleshooting. Three features of this design patterns are worth noting: First, it addresses troubleshooting at a level that guides assessment design across many domains for which its undergirding psychological perspective is appropriate. Second, it guides designers' thinking through categories that help ensure that a coherent argument results. Finally, it focuses on the nature of troubleshooting rather than on particular forms of evidence. Thus, the features of the design pattern conceptually unify assessments that would use different task types for distinct purposes or in particular contexts. Other examples of design patterns that stress interactive capabilities, whether in real-world or simulated environments, include experimental and observational investigation (Liu, et al., 2010; Mislevy, et al., 2009); systems thinking (Cheng, et al., 2010); and a suite of design patterns for model-based reasoning that include model formation, use, revision, and inquiry cycles (Mislevy, Riconscente, & Rutstein, 2009).

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). 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 Effect of 21st Century Technology on Domain Analysis and Domain Modeling

Not only has the current digital revolution changed the content of current constructs, it also affects how we may track and make sense of them. Specifically, technology has affected the practices of domain analysis and modeling by 1) changing the types of models of capabilities, environments, and performances we are likely to create; 2) providing new tools with which to analyze the domain; 3) allowing for the creation of digital representations; 4) providing means of collaboration around these representations; and 5) allowing for easier searching of these representations.

Writing from an economic-anthropological perspective, Perez (2003) has argued that technological revolutions not only change what we do but also alter the central metaphors that drive social discourse. For example, during the technological revolution of mass production (starting approximately 1900) the core technologies sought efficiency through decomposition of work and the use of hierarchical relationships. This led to educational distribution models based on mass production and organizational models in academia, business, and government centered on hierarchy. Following the more recent computer network revolution (starting with the use of the World Wide Web circa 1994), collaborative and computational networks have become a central metaphor in modern thought and discourse. Preceding this emphasis, many gains in experimental psychology were made through the information-processing revolution in psychology that arose from the mind as a computer metaphor (Anderson, 2009; Gardner, 1987). The new network metaphor has led to dramatic growth in the use of network representations (Barabasi, 2003) and analysis in the social and educational sciences (Freeman, 2004; Nooy, Mrvar, & Batagelj, 2005; Wasserman & Faust, 1994). These new metaphors affect our understanding of what it means to be proficient in a domain; how people become proficient; how we assess their developing capabilities; and consequently the types of models we are likely to construct when representing a domain.

The growth of digital artifacts can accelerate the analysis and communication of domain understanding. For example, prior to the recent digital revolution, updates to standards by academic groups would have to be communicated by physical mail and verbal communication based on close personal and professional relationships. Access to new professional standards may have taken years—as cycles of professional face-to-face meetings would serve as a core distribution methodology. Today instant communication via listserves, websites, and RSS feeds allows for rapid promulgation of new information. This will likely lead to a dramatic increase in the rate and volume of research in the years ahead, thereby increasing the rate change of the analysis and documentation of domains.

In addition to the more established technologies mentioned above, newer interactive web technologies could encourage input and embellishment from large groups of interested parties, following models such as wiki-nomics (Tapscott & Williams, 2008) or collective intelligence (e.g., Segaran, 2007). In such arrangements, (Wikipedia being the best known but only a single variation) technology is leveraged to allow for collaborative input and community evolution and resolution. These models have interesting economic dynamics (Benkler, 2007) that may have special advantages in certain highly collaborative educational environments where publishing was previously a bottleneck to the dissemination of knowledge as well as to the collaborative construction of new knowledge.

One method of dealing with the growing profusion of curricular and assessment resources tied to models of a domain (e.g., the plethora of state standards) is the use of detailed tagging associated with semantic web technologies. These tagging technologies support the machine-based search and aggregation of relevant information. They also support the machine-based inference regarding associations implicit in the organization of the data. This means that not all relationships need to be made explicit and that search is more forgivable and flexible, which can be incorporated in new technologies driven by algorithmic requirements as well as to human search. Such systems can be seen, for example, in the Achievement Standards Network (http://www.jesandco.org/). In sum, 21st century technology has and will continue to impact the types of analyses and models we create and the methods we use to create, display, and analyze them.

The Conceptual Assessment Framework

Domain analysis and domain modeling serve as core inputs to what has been traditionally considered assessment activity as described in ECD as a series of conceptual models called the Conceptual Assessment Framework (CAF). It is in the CAF that the domain information is combined with information regarding particular goals, constraints, and logics to create a blueprint for an assessment.

Assessment design activity can be thought of as a series of questions such as: "What are we measuring?", "How do we want to organize the world to collect evidence for the measurement?", and “What are the conceptual linkages between observable evidence and abstract inferences?" Whereas domain modeling addressed these questions as integrated elements of an assessment argument, the CAF expresses answers in terms of what amount to specifications for the machinery through which the assessment is instantiated. The CAF is comprised of a number of pieces called models that are composed of objects, specifications, and processes to this end. Objects and specifications provide the blueprint for the operational aspects of work, including the (a) joint creation of assessments, tasks, and statistical models; (b) delivery and operation of the assessment; and (c) analysis of data fed back from the field. Implementing these objects and coordinating their interactions in terms of the four-process delivery system (described in an upcoming section) brings the assessment to life. While domain modeling emphasized the interconnections among aspects of peoples' capabilities, situations, and behaviors, the CAF capitalizes on the separability of the objects that are used to instantiate an assessment. This becomes important in view of 21st century technology, as the models and their components can themselves be rendered in digital form. They are then amenable to assisted and automated methods of generation, manipulation, operation, and assembly (Mislevy, et al., 2010).

Figure 4 is a high-level schematic of the three central models in the CAF and their accompanying objects. . The specific elements they may contain in particular assessments are test specifications; item selection algorithms for adaptive testing; psychometric models; rubrics for raters or automated scoring routines; work product specifications; task models; and several others. The CAF contains the core of the evidentiary-reasoning argument—from task design to observations to scoring to inferences about students.

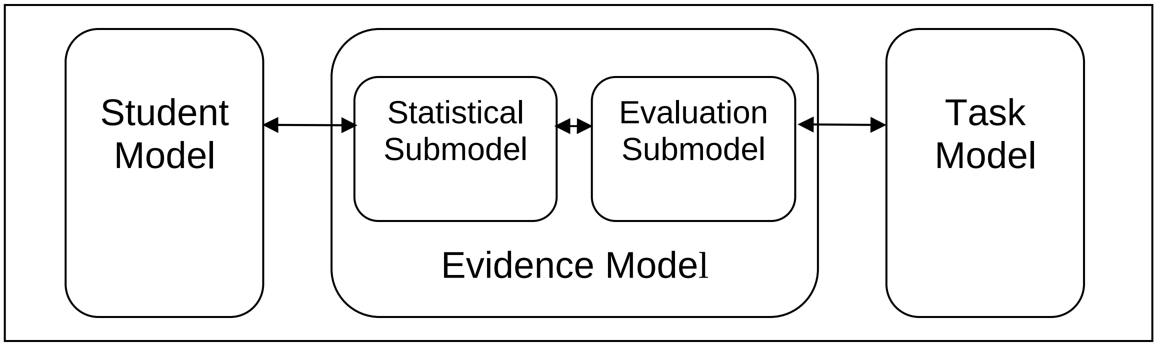


Figure 4. The central models of the Conceptual Assessment Framework.

The Student or Proficiency Model: What are we measuring?

The student model answers the question: *What complex of knowledge, skills, or attributes (KSAs) should be assessed?* A student model specifies a relevant configuration of the set of infinite configurations of skills and knowledge real students possess, as seen from some perspective about skill and knowledge in the domain. These are the terms from which we want to determine evaluations, make decisions, or plan instruction. The ECD literature sometimes refers to the elements of these models as Student Model Variables. As part of proficiency models following Williamson, Mislevy, & Almond (2004), we prefer a broader term and call them Proficiency Model Variables. The nature and number of proficiency model variables in an assessment depend on its purpose. A single variable that characterizes overall proficiency might suffice in an assessment meant only to support a pass/fail decision or a broad characterization of progress at the end of a chapter. However, a more complex structure would be required to report proficiencies on multiple proficiency model variables (e.g., a networking assessment used to assess students' routing, configuration, and troubleshooting). A complex structure might also be expected to provide diagnostic levels; make instructional decisions; or even change the situation in a game or simulation-based assessment (such as suddenly introduce a complication for a medical student that is performing well in a computerized patient management case). In this way, a proficiency model is likely to be a subset of the entities and relationships documented in the domain analysis stage, selected to align with a particular assessment goal and operationalized by the variables in a measurement model. Technically, the variables in a proficiency model are unobservable (latent) variables, such as those in psychometric models such as item response theory, latent class models, cognitive diagnosis models, and Bayesian inference networks. We will provide an example of the last of these shortly.

Evidence Models: How are we measuring it?

After the key KSA's are instantiated in proficiency-model variables, evidence models are used to identify the behaviors or performances that reveal these constructs and their relationship to them. An evidence model embodies the argument about why and how our observations in a given task situation constitutes evidence about student model variables.

The evidence model is composed of two parts: The evaluation submodel answers the question: What rules and procedures do we use to identify characteristics of work products as specific numeric or symbolic values for summarization and reporting? The outputs of these rules are called observable variables. The statistical submodel answers the question: With what weights and through what mechanisms do we want to combine information from performances, as summarized by values of these observables, to update our belief about proficiency model variables and at the same time understand and communicate our uncertainty about those values?

The evaluation submodel is represented in its most simple and familiar forms by responses to multiple-choice items and raters' evaluations of open-ended responses. Both 21st century skills and technology are revolutionizing evaluation in assessment. We have noted that critical aspects of expertise are manifest in interactions with people and situations, such as apprehending, constructing and reasoning through representations that are adaptive and evolving. An example would entail making choices and taking actions that create new situations. No longer is it simply a matter of a crisply defined task created solely by the assessor and a clearly separated response created solely by the examinee. Rather, the examinee’s moves continuously create new situations that in turn engender further moves and are often unique to every examinee that experiences the assessment. Qualities of the moves themselves— such as fluency, appropriateness, and effectiveness— are now targets of evaluation, as well as final products. A particular challenge is that making sense of an examinee’s actions requires understanding key features of the situations as they evolve.

Such types of assessments and evaluations were rare and costly when the performances were limited to live situations and evaluation was limited to human raters. A digital task environment, however, opens the door to automated data collection and evaluation. Technology in and of itself cannot determine what actions are important to capture, what to notice about them, and how to make sense of it; data is not the same thing as evidence. Thought processes are essential to crafting automated scoring in digital environments (Bennett & Bejar, 1998). Williamson, Mislevy, and Bejar (2006) provide an in-depth discussion of automated methods for evaluating complex performances, from the perspective of ECD.

The statistical submodel has been the sharp focus of the psychometric community for over 100 years. Approaches include classical test theory (Gulliksen, 1950/1987; Lord & Novick, 1968); generalizability theory (Cronbach, et al., 1972; Brennan, 2001); structural equation modeling (Kline, 2010); cognitive diagnosis models (Nichols, Chipman, & Brennan, 1995; Rupp, Templin, & Henson, 2010); item response theory (De Boeck & Wilson, 2004; Lord, 1980); and Bayesian Inference Networks (BNs) (Mislevy, 1994; Mislevy et al., 2003). Although all of these approaches are compatible with the ECD framework, Bayesian Inference Networks have been highlighted because their extensibility and graphical underpinning (both visually and computationally) align well with the central logic of ECD.

The networks are named because they support the application of Bayes’ theorem across complex networks by structuring the appropriate computations (Lauritzen & Spiegelhalter, 1988; Pearl, 1988). BNs properly and efficiently quantify and propagate the evidentiary import of observed data on unknown entities, thereby facilitating evidentiary reasoning under uncertainty as is warranted in psychometric and related applications (Almond, DiBello, Moulder, & Zapata-Rivera, 2007; Levy & Mislevy, 2004; Reye, 2004; Spiegelhalter, et al., 1993). In assessment, a BN is constructed by modeling performance on tasks (as summarized by the observable variables from the evaluation submodel) and as dependent on student capabilities (as summarized by the latent variables that comprise the proficiency model).

Once data are collected, values for the observables are entered into the network and the distributions of the remaining variables are updated. This constitutes evidence accumulation in assessment, where the evidentiary impact of the observed data yields posterior distributions for unknown proficiency model variables (Mislevy & Gitomer, 1996).

BNs are a powerful statistical modeling approach that offers a number of advantages for evidence accumulation in assessment, particularly for innovative and complex assessment environments. Like other statistical modeling approaches to assessment, they can support modeling of tasks in terms of psychometric features (e.g., difficulty, discrimination, etc.) and probabilistic inferences about students. Two particular advantages allow BNs to operate on the cutting edge of evidence accumulation. The first is BN’s flexibility for handling a variety of complexities that pose challenges to other statistical modeling approaches to evidence accumulation, which include conditional dependence among observations, relationships between task features and students’ capabilities, and multidimensional latent variable models where performance on tasks depends on multiple and distinct (though possibly related) skills or other aspects of proficiency. This flexibility allows the analyst to specify a wide range of relationships reflecting theories of task performance and skill acquisition (Reye, 2004; VanLehn, 2008), including situations with multiple, confounded, and serially dependent observations. Yet the second advantage is more important in interactive digital environments. Using BNs to propagate the inferential force of observed data allows for the construction and piecing together of BN fragments in light of the features of an evolving situation and the examinee’s actions up to that point. This capability uniquely supports dynamic and evolving assessment situations, from adaptive testing (Almond & Mislevy, 1999) to intelligent tutoring systems (VanLehn, 2008) and on-the-fly continuous assessment in multiplayer games (Shute, Hansen, & Almond, 2008).

Task Models: Where do we measure it?

Task Models answer the question: How do we structure the kinds of situations are necessary in order to obtain the kinds of evidence we need for the evidence models? This includes the presentation material and the work products. Task Models also answer the questions: "What are the features of the tasks" and "How are those features related to the presentation material and work products?"

Many 21st century skills revolve around construction, communication, and interaction; they are enacted over time and across space, often in virtual environments, with cognitive and digital tools and representations. The capabilities for creating these environments for students during assessment have arrived; we are able to build complex simulation environments that mirror or extend the real world. Yet this is not the same as saying that we know how to leverage these environments to carry out assessment efficiently and validly. Building and making explicit assessment arguments in domain modeling goes a long way toward validity. Specifying the objects and processes to instantiate the arguments in reusable, interoperable objects that match the assessment argument goes a long way toward efficiency. In designing simulation- and game-based assessment, we want to build features that require the targeted capabilities and provide affordances for students to enact their thinking. Moreover, we seek to do this with code that we can re-use multiple times in many combinations with customizable surface characteristics. In the Cisco Networking Academy, local instructors as well as test developers are provided a designer interface that allows them to easily create and share simulation tasks in the Packet Tracer environment (described in more detail below). They use standard tools, representations, and affordances; and automated scoring routines to identify features of final configurations that are produced automatically for these tasks (Frezzo et al. 2009).

An important constituent of ECD task models is task model variables. These are features that task authors use to help structure their work in several ways, including effectively defining proficiency model variables; controlling difficulty; assembling tests to meet target specifications; and focusing the evidentiary value of a situation on particular aspects of skill (Mislevy, Steinberg, & Almond, 2003). The impact of different task features can be incorporated into the statistical submodels to improve the inferences about students in terms of proficiency model variables and/or to yield information that can be incorporated into assessment assembly to improve task construction and selection (De Boeck & Wilson, 2004).

The Effect of 21st Century Technology on the CAF

Psychometric advances in the statistical aspects of evidence models in the 20th century were largely predicated on logical and statistical independence of tasks. In addition, the evaluation submodels were largely predicated on fixed response scoring and delivery models were restrictive in what could be provided. These limitations, together with a behavioral understanding of human activity, often contributed to assessments constructed in a highly atomized manner, with little resemblance to the real-world tasks through which inferences about performance were being made.

The growing presence of digital experiences in everyday life as well as the emergence of interactive media and simulation for education and entertainment both foreshadow the merging of digital tasks created for non-assessment uses and become the basis for assessment inference moving forward. Indeed many common interfaces are equipped with built-in assessment systems as part of the daily experience. For instance, Microsoft Word has an embedded evidence model that observes a user's typing and makes inferences about intended and actual spelling, and may automatically make corrections to the text or provide suggestions for alternate activity. This is a type of on-the-fly embedded assessment encountered as a non-assessment activity in day to day experience. Similarly, as we will discuss in the following section, educational games contain many of the elements of evaluation of a work product and presentation of new tasks present in assessment. The difference between the assessment performance and the learning or work-based performance can be reduced with digital assessment, compared to task design in physical modes.

From our perspective, these are examples of the fusion of assessment driven tasks and daily-life-driven-tasks enabled by the recording, storage, and manipulation of digital information. Recording, storage and manipulation requirements have historically been requirements of assessment inference but they are becoming general standards for many activities in our digital lives. This opens up the opportunity for what Behrens, Frezzo, Mislevy, Kroopnick, and Wise (2008) called ubiquitous unobtrusive assessment and what Shute, Hansen, and Almond (2008) call stealth assessment. As the world continues to evolve along the line of increasingly digital interactions (Bell & Gemmell, 2009), assessment designers should continue to ask: Does new information need to be collected or is it already occurring in the environment."

The key to utilizing voluminous digital information is recognizing, at a level of generality higher than the particulars of situations and actions, the patterns that signal cognitively important patterns—features of the situation, as well as features of the action. In troubleshooting, for example, what are essential features of the countless troubleshooting situations that admits to space-splitting and what are essential features of action sequences in these situations that suggest that this is what the examinee has done (see Steinberg & Gitomer, 1996, for answers to these questions in the context of troubleshooting aircraft hydraulics systems)?

Assessment Implementation and Delivery: The Four Process Model

The delivery system used in the Networking Academy Program follows the Four Process architecture suggested in the ECD framework described in Almond, Steinberg, & Mislevy (2002). As illustrated in Figure 5, this view divides the delivery aspects of assessment into four core components. 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; and an intelligent tutoring system can jump out to instructional or practice models. The form of the logic, the processes, and messages have been specified out in the CAF models, which in turn have been developed to instantiate the assessment argument in domain analysis. Almond, Steinberg, & Mislevy, (2002) delineate the CAF relationships between the CAF models and the delivery system processes. In this way, we see the usually-hidden role that the pieces of machinery in an assessment play in reasoning.

Activity or Task selection revolves around the choices that will be presented to the examinee. This may be as simple as a rule to show the next question or may be based on a complex mathematical model of examinee knowledge and dispositional states. The next process is called presentation, which considers the presentation of information to the examinee and the acquisition of information from the examinee.

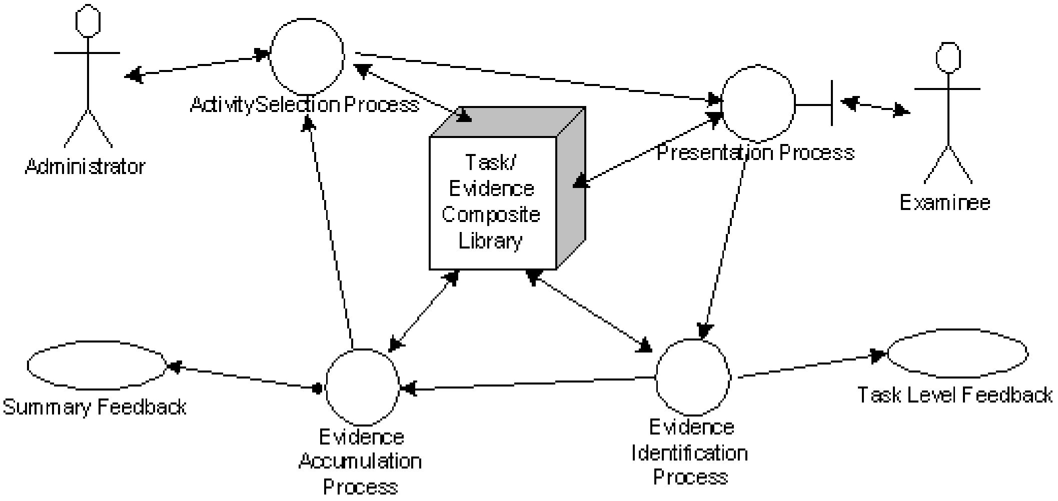

Administrator
ActivitySelection Process
Task/ Evidence Composite Library
Presentation Process
Examinee
Summary Feedback
Evidence Accumulation Process
Evidence Identification Process
Task Level Feedback

Figure 5. Processes in the assessment cycle (from Mislevy, almond, & Lukas, 2004).

The result of the presentation process is a work product that must be examined and scored using rules derived from the evaluation submodel in the CAF. This process is called response processing or evidence identification and constitutes the third of the four processes. The ECD literature discusses the idea of looking for features in the work product and characterizing the work product in terms of one or more observable variables. It is important to note that work products can include process information— such as log files, captured screen actions, or video files of a performance). These observable variables characterize some aspect of the work, such as correctness, efficiency, or fluency. They are also characterized by more specific features like whether a sequence of actions is consistent with space-splitting or whether a router appropriately passes the messages it should allow to a particular computer and blocks the ones it should not allow. Most multiple choice questions are scored on a single observable of correctness. However, complexity is relatively easy to add— even from this simple work product. For example, questions can be written with scoring rules, such that if a student chooses option A, he or she may receive 1 point each for correctness and efficiency. Option B may represent an answer that receives 1 point for correctness and 0 points for efficiency. The fourth process, evidence accumulation, combines information from all the observables and updates ability estimates in the student model through the statistical submodels in the CAF. If the assessment delivery has multiple phases, the activity selection process will decide what to do next.

Twenty-first century technology influences how assessments are delivered as conceptualized in the four-process model. Advances in computer-based semantic analysis and pattern matching allow for more complex evidence identification procedures. In both digital and physical systems, pattern recognition is used to map aspects of the work product to symbols that represent quality or quantity of relevant aspects of the work product (i.e., values of observable variables). In physical scoring systems, the need for mechanical simplicity in scoring often drives backward up the ECD chain to constrain tasks to simplified formats consistent with simplified scoring, thereby constraining the kinds of evidence that can be obtained. In digital systems, we can program rules and algorithms to identify and process a wider variety of types of work products and apply scoring rules more flexibly.

Concomitant with this change in evidence identification are corresponding changes in evidence accumulation. The previous discussion of Bayesian Networks as statistical submodels highlighted that BNs help us define how we gather evidence regarding proficiency model variables. They also are a sophisticated method of evidence accumulation that complements advances in computing technology to create adaptive learning and assessment tools. When a person completes a task in an adaptive assessment or intelligent tutoring situation, the resulting values for the observables that constitute evidence about proficiency are used to update the BN. This updates the estimates of the proficiency model variables of interest, including estimates of our uncertainty. These updated estimates can then be used to select the next task to present to the student. See Mislevy & Gitomer (1996), VanLehn (2008), and Shute et al. (2008) for detailed descriptions of applications.

An Application of ECD Using 21st Century Technology

One goal of the Cisco Networking Academy is to provide instructional support in order for students to become proficient networking professionals. Students need both a conceptual understanding of computer networking and the skills to apply this knowledge to real situations. Thus, hands-on practice and assessment on real equipment is an important component of the curricula. However, we also want to provide students with an opportunity to practice outside of class, explore in a low-risk environment, and build complex networks with more equipment than an average classroom has available.

To address these needs, Cisco has developed Packet Tracer (PT), a computer program that provides simulation, visualization, authoring, and assessment to support the teaching and learning of complex networking concepts (Frezzo, et al, 2010). The PT software supports the authoring and distribution of network micro-worlds whose logic and activity are highly simulated at several levels of complexity while also providing interfaces to support explanatory and assessment purposes. Numerous PT activities are pre-built into the current curricula (upward of 150 in some courses); instructors and students can construct their own activities for it; and students can explore problems on their own. For the purposes of this presentation we will focus on the affordances of PT that make it an effective assessment platform. Assessments in the Networking Academy fall into the categories of student initiated or instructor initiated. Student initiated assessments are primarily embedded in the curriculum and include quizzes, interactive activities, and PT challenge labs. These interactions provide feedback to the student to help their learning and use a wide array of technologies, including multiple-choice questions (in the quizzes) and complex simulations (in the challenge labs). Until recently, instructor initiated assessments consisted either of hands-on-exams with real networking equipment or multiple-choice exams in the online assessment system. As of 2010, this system additionally provides users with a variety of simulation-based end-of-chapter and end-or-course feedback and grading events. The assessment activity described below is called the Packet Tracer Skills Based Assessment (PT SBAs). It integrates the flexibility and detailed feedback provided in the student-initiated assessments with the data collection, detailed reporting, and grade-book integration available in the more traditional instructor-initiated assessments. Examination of how the four process model described above is implemented with the PT demonstrates how technology makes these assessments possible.

Task Selection

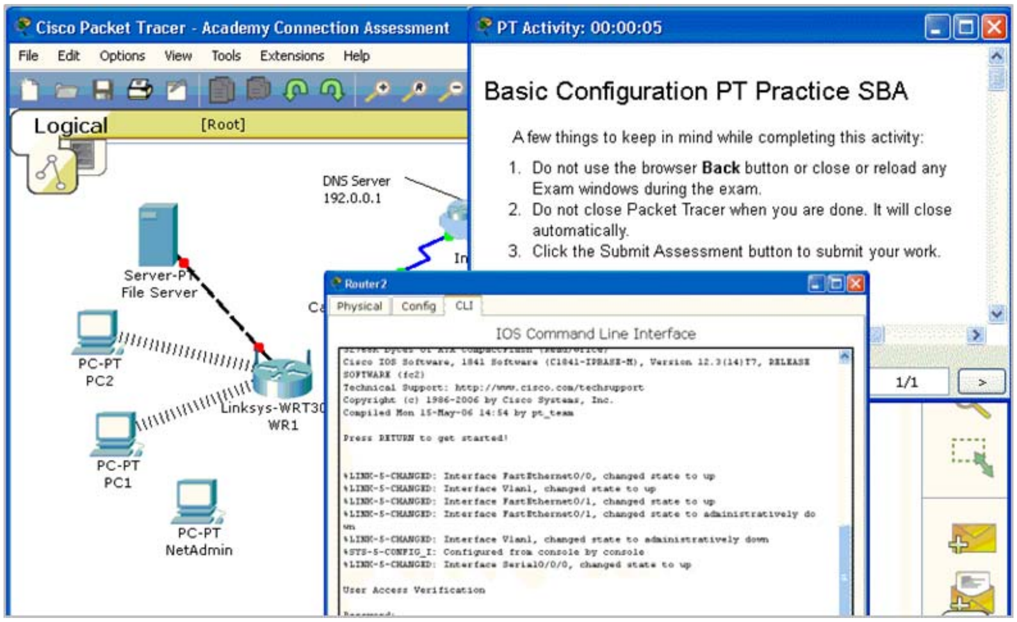
Task Selection is the process that is least automated in the current PT SBAs. Each assessment consists of one extensive network configuration or troubleshooting activity that may require up to 1.5 hours of work across multiple sub-tasks. Access to the assessment is associated with a particular curricular unit and it may be re-accessed repeatedly based on instructor authorization. PT assumes a task is selected for administration (by loading a network and task file) and that PT will act over the remaining three processes. In the future, it is possible that smaller tasks could be created and selected based on results of the previous tasks.

Presentation

The rich interface and drag-and-drop interaction of PT is a hallmark of the software. It provides a deep (though imperfect) simulation of a broad range of networking devices and networking protocols, including rich features set around the Cisco IOS. Instructions for tasks can be presented through HTML formatted text boxes that can be pre-authored and locked by any user. In this way, PT is not simply a simulation tool but actually a micro-world authoring and interaction tool with instructional and assessment affordances. One important differentiator between PT and many other instructional environments is its variable manager feature that allows the template creation of networks and the generation of random version of the network based on ranges of values. Values can be generated by random selection from lists or numeric ranges in both the textboxes that describe the task or activity, as well as in values of much of the network data (e.g. IP addresses). This allows for the development of large number of practice examples or isomorphic tasks to be generated at run time (Frezzo et al. 2010).

The micro-world environment in PT simulates a broad range of devices and networking protocols including a wide range of Personal Computer (PC) facilities covering communication cards, power functionality, web browsers, operating system configurations etc. The particular devices, configurations, and problem states are determined by task authors guided by design patterns, in order to address whatever proficiencies are targeted by the chapter, the course, or the instructional objective. When icons of the devices are touched in the simulator, more detailed pictures are presented with which the student can interact. A broad range of networking devices are simulated include routers, switches, internet-based phones and video devices. Network management protocols are simulated from simple formats that may be used in a home to complex protocols used to manage portions of the internet.

Because the program provides assessment support to schools in over 160 countries with varying network bandwidth, there is a computational need to balance the size of a robust simulation engine with the need for a relatively small assessment definition package from the central WWW server. To accomplish this, the system is constructed to have the PT simulation program installed on the user's desktop before the assessment is initiated. This is usually non-problematic as the systems are generally pre-installed to support the curricular use of PT. At assessment run time, the assessment system sends the PT instance the micro-world definition file (a .pka file) which includes all the information necessary for the micro-world definition, and assessment scoring instructions (Figure 6). When the activity is submitted, the entire student network is relayed back to the server along with detailed log files and the results of the evidence identification work that was accomplished in the PT software.



Basic Configuration PT Practice SBA

A few things to keep in mind while completing this activity:

1. Do not use the browser **Back** button or close or reload any Exam windows during the exam.
2. Do not close Packet Tracer when you are done. It will close automatically.
3. Click the Submit Assessment button to submit your work.

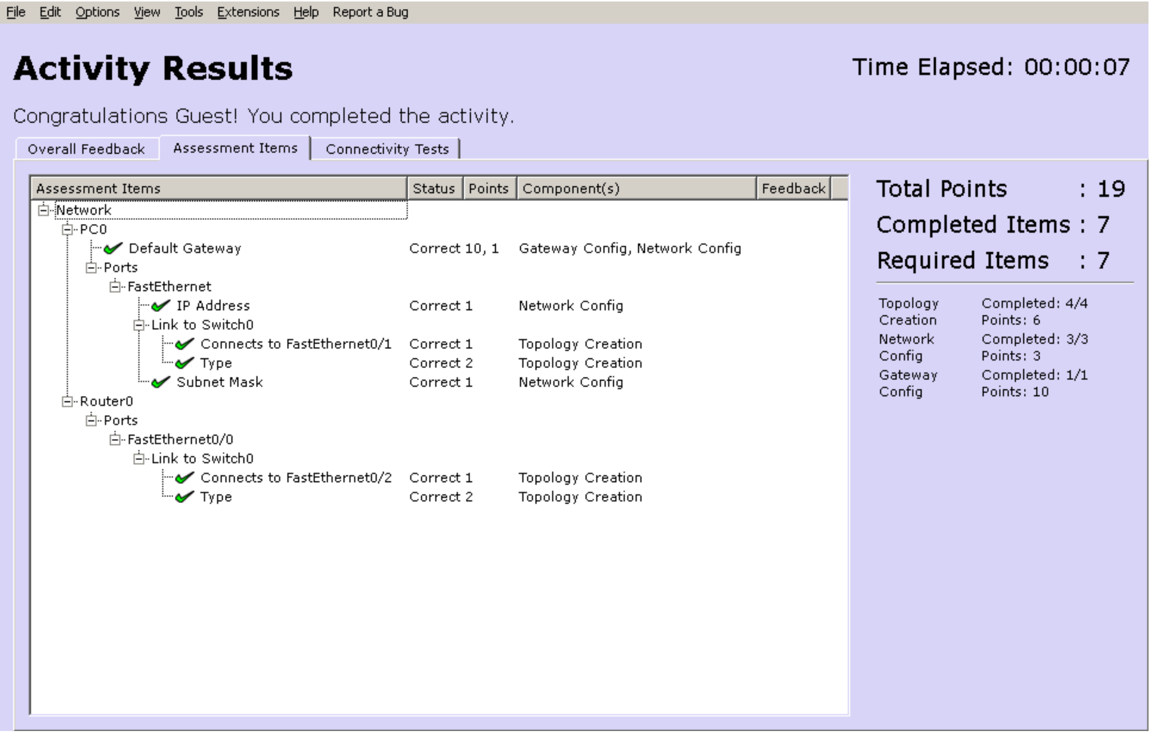
IOS Command Line Interface

```
Cisco IOS Software, 1841 Software (C1841-IPBASE-M), Version 12.3(14)T7, RELEASE
SOFTWARE (fc2)
Technical Support: http://www.cisco.com/techsupport
Copyright (c) 1986-2006 by Cisco Systems, Inc.
Compiled Mon 15-May-06 14:54 by pt_team

Press RETURN to get started!


%LINK-5-CHANGED: Interface FastEthernet0/0, changed state to up
%LINK-5-CHANGED: Interface Vlan1, changed state to up
%LINK-5-CHANGED: Interface FastEthernet0/1, changed state to up
%LINK-5-CHANGED: Interface FastEthernet0/1, changed state to administratively do
wn
%LINK-5-CHANGED: Interface Vlan1, changed state to administratively down
%SYS-5-CONFIG_I: Configured from console by console
%LINK-5-CHANGED: Interface Serial0/0/0, changed state to up

User Access Verification
```

Figure 6. Screenshot of a Packet Tracer skills-based assessment.

Evidence Identification/Task Scoring

The approach taken in PT to assist with the authoring of scoring rules needed to be easy to use, comprehensive, and consistent with the flexibility of the ECD model. To accomplish this, PT provides an Activity Wizard that allows the construction of an Initial Network and an Answer Network. The Initial Network is the starting state of the network micro-world. As depicted in the tree on the left side of Figure 7, the Answer Network is the matching key created by the states of comparison network configured for scoring by providing the author with a comprehensive list of network states. The original state of the Answer Network provides a list of potential work product features of interest to the assessment designer. These are low-level features, including how particular aspects of the device are configured; whether a cable is plugged in; and in what ways traffic on the network is occurring. By considering the purpose of the assessment and the relationship between work product features and signs of proficiency, the designer checks the boxes of all the features about which they would like to obtain data.



Activity Results

Time Elapsed: 00:00:07

Congratulations Guest! You completed the activity.

Assessment Items	Status	Points	Component(s)	Feedback
Network				
PC0				
✓ Default Gateway	Correct	10, 1	Gateway Config, Network Config	
Ports				
FastEthernet				
✓ IP Address	Correct	1	Network Config	
Link to Switch0				
✓ Connects to FastEthernet0/1	Correct	1	Topology Creation	
✓ Type	Correct	2	Topology Creation	
✓ Subnet Mask	Correct	1	Network Config	
Router0				
Ports				
FastEthernet0/0				
Link to Switch0				
✓ Connects to FastEthernet0/2	Correct	1	Topology Creation	
✓ Type	Correct	2	Topology Creation	

Total Points : 19
Completed Items : 7
Required Items : 7

Topology Creation	Completed: 4/4 Points: 6
Network Config	Completed: 3/3 Points: 3
Gateway Config	Completed: 1/1 Points: 10

Figure 7. Packet Tracer results screen showing features of Answer Network (left), observable values, component loading values, and components.

After indicating the features from which to create observable variables, the assessment designer edits the answer network values to create implied scoring rules. At time of task completion, the target features of the learner network are compared pair-wise with the corresponding features of the answer network. The comparisons used to generate observable variables may look for exact matches (answer network cable plugged in, examinee network cable plugged in) or allow for evaluation of range of values (answer IP address between xx.xx.0 and xx.xx.100, examine network IP address xx.xx.50.). This pattern matching is made flexible by both the use of the variable manager and the use of regular expressions in the evaluation clauses. Network functionality tests can also be authored in addition to answer network end-state tests. In other words, the scoring and the creation of algorithms necessary for automated scoring are automated. They are driven by simple choices that the author makes in regards to which aspects of the work product need evidence. Areas of proficiency reporting associated with student model variables can also be specified and associated with observable summation, as shown on the right side of Figure 7.

In the most basic interface, the Activity Wizard provides ease of use and structure to support basic assessment authoring; in addition, PT also gives instructors and other assessment designers direct access to many of the micro-world variables using a comprehensive macro language. By providing this custom programming layer, more detailed observable definition and combination schemes can be created from the network values. Recent versions of PT have also included the availability of a macro-language for directly accessing micro-world states, translating them into observable values and combining the lower level observables into higher order combinations.

While log files are not currently evaluated by the PT scoring engine (see future directions below), they are captured and available for further review by students and instructors through the online grade book. This allows instructors and students to review the work of the student in detail and make custom decisions or discussions enabled by the assessment infrastructure. The files can also be exported to automated systems for analysis and evaluation. Another feature of Evidence Identification in the ECD framework is that it is the minimum requirement for providing feedback. After some aspects of the work products have been characterized in terms of observable variables, these variables can be used to trigger feedback to learners. Moreover, PT allows the authoring and reporting of observable level feedback. That is to say, different strings can be presented to the learner depending on whether a specific work product feature, or combination of features, is present or not. In the results window provided to the student (as seen in Figure 7), the values of the observable variables are communicated with the labels of correct and incorrect or any other value-specific string desired. For instance, text identifying specific strategies or potential interpretations can be authored as well.

Evidence Accumulation

While PT was not designed originally as an assessment and measurement tool, it has important features that support the facilitation of linkages from observables to variables called components, which serve as proficiency model variables in the ECD framework. For each observable variable, PT allows the specification of multiple components to be associated with the observable variable and allows the specification of differential weights. In the ECD model, this is described as a multiple-observable/multiple-proficiency model variable architecture. An important concept is that the different observable variables can provide information in multiple dimensions; for example, establishing communication between two routers can depend on a student's understanding on two dimensions: IP Addressing and Connectivity. Accordingly, it is important to conceptualize the observable variables not simply as identifiers of correctness, but rather as a piece of information about a feature of performance that provides information for one or more proficiency model variables. In many traditional assessment systems, each task generates one observable and updates exactly one proficiency model variable—a simplifying assumption at odds with the integrated use of multiple aspects of knowledge and skill that characterizes most problem-solving in the real world. The primary limitation of the PT software in this area is that loadings between the observables and the components (proficiency estimates) are limited to standard algebraic functions found in common computer languages (which are thereby passed to the macro language). However, because PT's architecture allows communication of information to external systems, one possibility is that future versions could allow probabilistic updating using more complex algorithms including BN methods (Wainer, Dorans, Flaugher, Green & Mislevy, 2000).

As of fall 2010, after an extensive beta test period, the assessment is in the full production systems and approximately 2,000 PT SBA are being delivered each week across eight courses. The collection observable values, logs, and final networks is providing a growing corpus of data with which to understand the variations in performance, missteps and expertise with which we plan to refine our scoring rules and reporting features. A survey of 141 instructors during the beta test period indicated an average satisfaction rating of 4. 5 on a scale anchored by 1=Very Dissatisfied; 3=Neutral; and 5=Very Satisfied. A survey of 916 students indicated an average satisfaction score of 3.9 with students taking the exam at home showing significantly lower satisfaction than those taking the exam at school. This is likely a side effect of configuration and network dependency issues that are idiosyncratic to home networks. Analysis of the patterns of observables for each exam suggested patterns of performance consistent with expectation.

The feasibility of such a complex system was made possible because each system was designed with ECD conceptualization and orientation toward the four process delivery model. The scoring back-end of PT provides the flexibility and support for task generation, evidence identification and evidence accumulation, as well as the ability to communicate to other systems that might need to augment or replace one of the four processes. Integration with the core multiple choice assessment system and corresponding grade book was possible because that system had been written from a four process model perspective and included a four process model extension module that allows for robust extension of the system to performance based input systems such as PT (Behrens, Collison & DeMark, 2005). By creating systems under the view of common ECD architecture, we are able to integrate new innovations over time, and add technologies that naturally promote reuse, efficiency and extensibility.

Future directions, Current limitations

Despite the advances in the application of evidence-centered design to new assessment technologies, there remain both limitations and opportunities. This section will discuss three areas that present opportunity for future work and the Networking Academies' efforts in these areas.

Games and Embedded Assessment

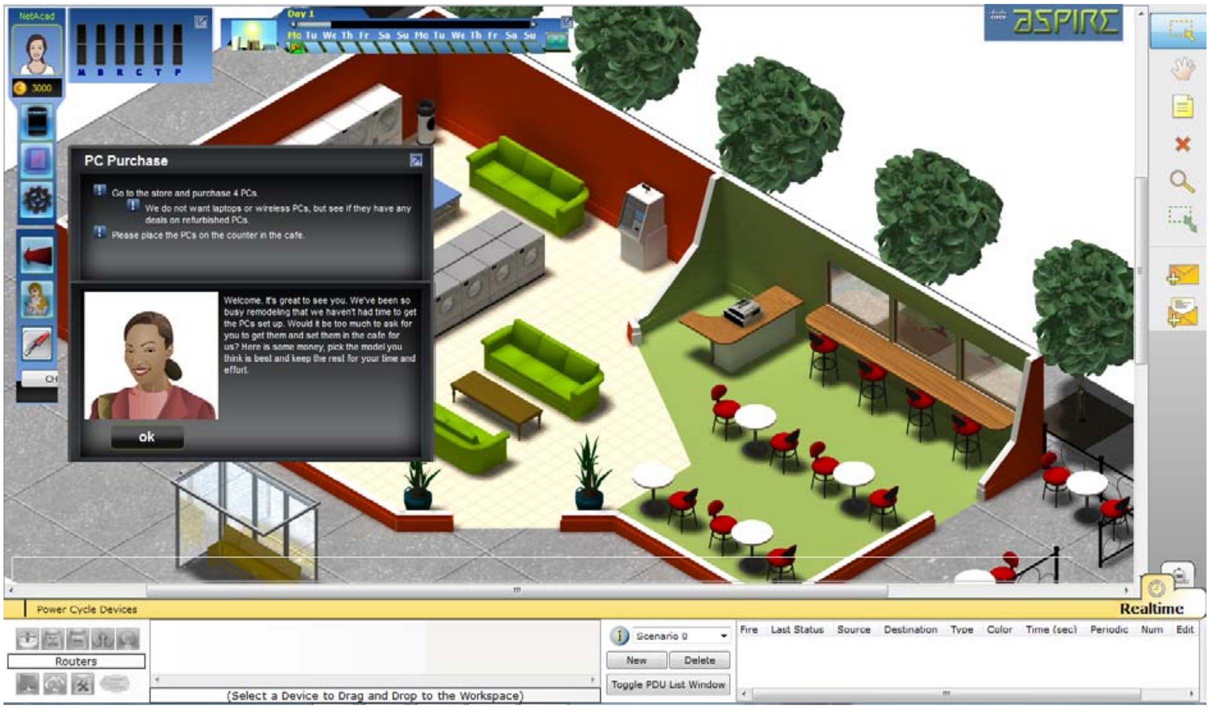
The Packet Tracer Skills Based Assessments described in the previous section are clearly defined and understood as classroom assessments for students. This understanding does not come from the structure of the software or tasks but rather the control, use, and implications of the activity (Frezzo et al., 2009). To provide high quality instructional and learning support, while avoiding the constraints and costs of high stakes testing, we next sought to create an assessment tool that would move us toward the goal of ubiquitous unobtrusive assessment (Behrens, et al., 2008). In other words, the goal was to, build the affordances of assessment (tracking and feedback) into the fabric of the student daily activity. While a number of approaches could be taken, we decided to extend the micro-world infrastructure and ECD-based scoring system in PT to create a complex game- like environment, thereby formalizing the work introduced in Behrens, et al. (2008).

Games are seen as attractive potential learning tools because they engage and immerse players in ways that traditional school content does not, providing the context needed to encourage application of learning (Gee, 2003). Games often involve spontaneous learning and demonstration of concepts through play (Clark et al, 2009), and they can elicit particular ways of thinking. Shaffer (2006) defines epistemic frames as the ways people in a given field decide what is worth knowing; agree on methods of adding new knowledge and standards of evidence; have concepts and representations to understand situations; and interact with each other and the world. When epistemic games are properly developed, they can engage people to think like doctors, scientists, or network engineers.

Behrens et al., (2008) argued that simulation-based games themselves contain many parallels to assessment. For instance, games and assessments both describe knowledge and skills in a quantifiable manner. Rules define what information is available and the constraints around solution paths. The ECD four-process model describes activity selection, presentation, response processing, and evidence accumulation in assessment and can also be applied to simulation game scenarios (Behrens, et al., 2009). Both assessment and simulation communities desire to create models of student (player) behavior and knowledge and often use similar tools (e.g., BNs) to do so.

Given the promise of games in the assessment sphere, the question then becomes how to make this potential a reality. Shute, Ventura, Bauer, & Zapata-Rivera (2009) explore embedding formative assessment within games. They advocate for the use of unobtrusive measures of performance that can be gathered while students maintain flow in the game; ultimately, this can help provide direct feedback on personal progress and/or modify the learning environment for the player. They introduce the term stealth assessment to describe embedded assessments so closely tied into the environment that they are invisible to the user. In the process of game play, students perform the very skills we would like to assess. We might capture their performance of these skills and consequently provide information about students' abilities. Shute et al. (2009) use the commercially available game Oblivion to demonstrate the use of ECD to assess creative problem solving. The paths that users take through the game (e.g., how they cross a river) serve as observable measures in task models that then inform evidence models and competency models. This allows for estimates of students' creative problem solving skills based on their game play.

The Cisco Networking Academy is in the early stages of experimenting with games as providers of assessment information. It recently released a networking and entrepreneurship game named Aspire (see Figure 8). The main idea of the game is that students are entrepreneurs who own small networking companies, and must make both business and technical decisions in the game. The Aspire system consists of a 21/2-D interface that allows navigation; interaction with characters in the game; decision making and interaction (sometimes in the form of multiple choice questions); and complex scenarios that combine numerous task requirements. This interface is integrated with the PT software which renders and simulates the computer and networking devices and systems as well as shows the ECD-based scoring architecture. This provides a high degree in design and analysis re-use between PT, the PT SBA and the Aspire game. We are in the early stages of analyzing and working with the data from what early student reactions suggest is an engaging, stimulating, and informative tool.



PC Purchase

- Go to the store and purchase 4 PCs.
  - We do not want laptops or wireless PCs, but see if they have any deals on refurbished PCs.
- Please place the PCs on the counter in the cafe.

Welcome. It's great to see you. We've been so busy remodeling that we haven't had time to get the PCs set up. Would it be too much to ask for you to get them and set them in the cafe for us? Here is some money, pick the model you think is best and keep the rest for your time and effort.

Figure 8. Screenshot of Aspire game.

Understanding Trace Data

The advent of computer-based simulations and gaming described above has brought the ability to capture highly detailed data as students progress through the environment. Data ranging in granularity all the way down to individual mouse clicks is available, which thus creates vast stores of information. The challenge lies in how to determine which data are useful and how to make use of this data in ways that will ultimately inform and improve student learning. These efforts fall primarily into the category of evidence identification. How do we take these work products and apply scoring rules that will provide meaningful information about students' knowledge, skills, and abilities?

In computer networking, one of the primary means of connecting networks is via programming routers and other network devices. In many contexts, these machines (which route data traffic such as your email or web page request) need to be programmed in order to be alerted of their location and the rules required for providing or denying access. This programming produces logs of commands and is one of the work products a PT SBA. We will discuss ways we have explored analyzing these streams of commands, and believe the same concepts could be applied if, instead of each data point being a command, it was a game location or mouse click. Of the many ways to think about this trace data, we will briefly discuss three: thinking about strings as words and sentences, thinking about strings as documents, and thinking about strings as neighbors.

Thinking about strings as words and sentences leads us to the field of statistical natural language processing (NLP; Manning & Schuetze, 1999). If we think of individual performances as text streams, it raises questions such as: Can we understand the relationships between different elements in the stream? Can we develop succinct descriptions of a performance? Can we extend these techniques to help us with more broadly use techniques such as clustering? DeMark and Behrens (2004) began this process with router logs. This work has continued with the data from the PT skills exams. NLP includes some common data techniques that have been helpful. For example, tokenization (or breaking the stream of data into meaningful segments) and creation of n-grams (groups of n tokens that occur together) allow us to identify the commands that occur together. Using tokenization, stemming, tagging, and other NLP techniques, we can examine command use in the entire sample, as well as within novices or experts, and tie this information back to the observables measured.

A second way to think about trace data is as texts in a corpus (Biber, Conrad & Reppen, 1998). This may allow us to understand variability in performances as a whole across individuals, describe similarity and dissimilarity between performances, cluster performances, and find the most informative dimensions of variation within and between performances. We can cluster individuals based on the patterns and sequences of commands (Gries, 2009). Ultimately, we believe this process will allow us to inform instruction by identifying and suggesting strategies that successful and less successful students employ.

A third way to think about trace data is as a network. In trace data, each location, click, or command is in a sequence before some events and after others. These could be thought of as neighbors, or members of a network, and the tools of social network analysis employed to investigate the relationships (cf. De Nooy, Mrvar & Batagelj, 2005; Stevens, Johnson, & Soller, 2005). In router logs, there are different probabilities of commands occurring near each other. We can use visualization tools to help us understand the networks. Figure 9 is a visual depiction of an entire exam's command sequence from one individual. Based on this work, we can begin determining things like the average distance from one command to another; identifying salient features of networks that differentiate experts and novices; and cluster individuals based on these network elements.

These three conceptualizations of trace data provide us with various lenses through which to understand large amounts student data gathered from computer-based performance assessments. In addition, future research may explore the use of neural network and support vector machine methodologies to examine this data, thinking about the data as observables. Consideration must also be given to automation processes and machine data learning given the large amounts of data under review. We believe there are numerous opportunities for research and advancement in this area.

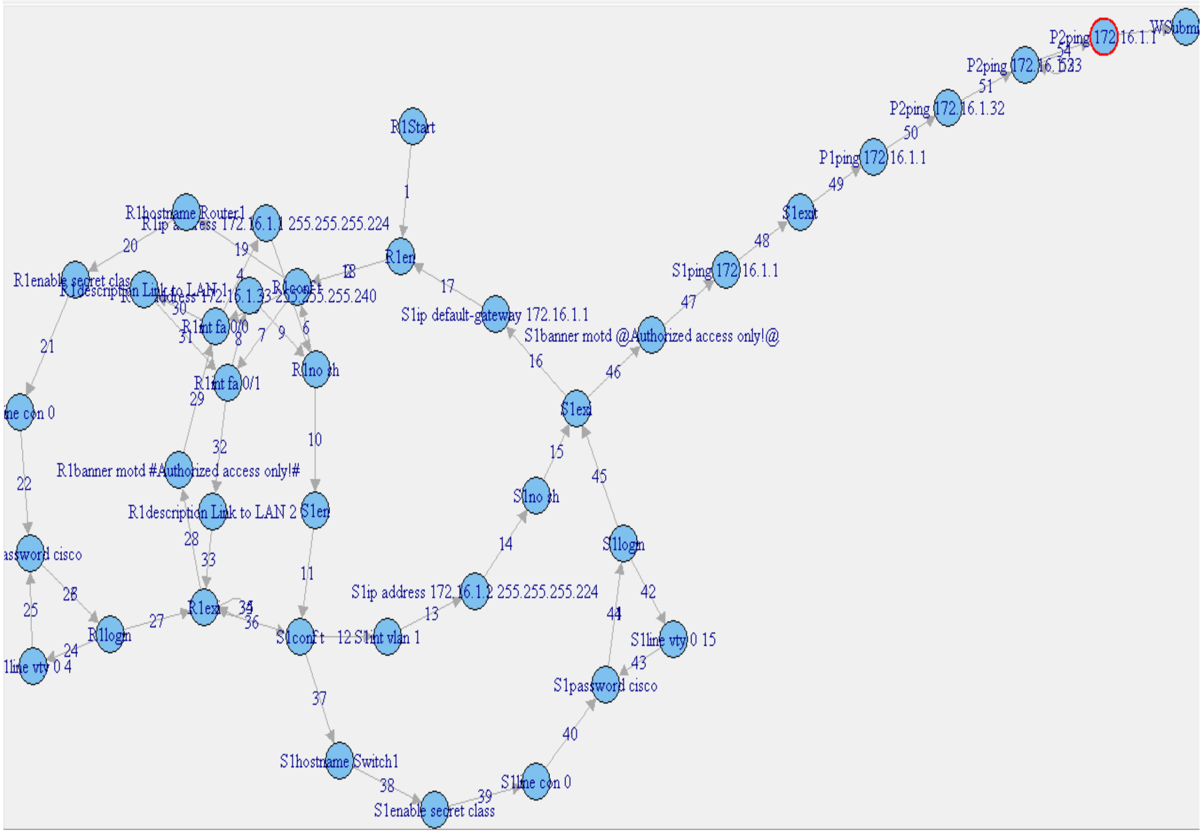


Figure 9. A visual depiction of the entire sequence of log commands for one student.

Curriculum-Assessment-Gaming Integration

Finally, a third area of opportunity we are exploring is the integration of information and data. Oftentimes curricula, assessment, and games are entirely separate projects, with no concerted effort to bring information from all sources to an understanding of student learning. Assessments are conducted by district, state, or national entities; curricula are produced and distributed by publishing companies; and games are produced by yet another set of companies or academic researchers. We would argue that to create a seamless flow of information about student learning, these three areas should be integrated. In that effort, we are pursuing efforts to integrate data obtained from students using the curriculum (e.g., practice PT activities); taking assessments (both formative and summative, performance and traditional); and playing games. We believe that with this integration we will be able to provide instructors and students with detailed feedback about their progress and make recommendations of resources, activities, and interactions that will further them along that progression while lowering the need to stop instruction for administratively driven assessment. This is consistent with the previous discussion of embedding assessment in the fabric of the digital world—in this case, the world of digital online learning.

The conceptualization of learning progressions has helped us with this integration process. Learning progressions are empirically grounded and testable hypotheses about how a student's understanding and ability to use knowledge and skills in a targeted area develop over time (Corcoran, Mosher, & Rogat, 2009). Learning progressions are being developed in the Networking Academy based on: the results of statistical analyses of millions of student exams taken over the life of the previous four-course curriculum; subject matter expert (SME) input; and the results of cognitive task analysis research into novice and expert performance in the curriculum domain (DeMark & Behrens, 2004; West et al., 2010). The learning progression analyses identify conceptual development of strands of increasing complexity/sophistication that we can then use to develop curricula, assessments, and games as well as bring together data from these sources in a meaningful way.

We believe that the BNs described above provide a way to model performance on learning progressions using data from different sources. West et al. (2010) provide an example of the use of BNs to analyze progress on a single learning progression using only data from multiple choice tests. Future work will undertake the challenges of utilizing data from multiple sources, modeling tasks that are influenced by multiple learning progressions, and modeling progression over time. The issue of modeling tasks that are influenced by multiple progressions is likely to be particularly important in the assessment of 21st century skills. In these assessments, nearly always some content expertise is needed to complete the question along with the 21st century skill of interest. For example, it is difficult to assess problem solving skills without a particular context for the assessment. Hence, it will be important to model both the level of the content learning progression and the 21st century skill progression of interest.

Concluding Observations

A fundamental advantage of designing assessments in an ECD framework is that it is flexible enough to accommodate the affordances of new technologies and the demand to measure new domains. These assessments also provide a unified framework, which describe current practice across a wide range of assessment activities. We have seen major advances in assessment practices because of the availability of 21st century technology. Furthermore, we are witnessing the beginning of additional changes and there are surely other unimaginable developments that await us. Similarly, the domains we want to measure will change as the demands of jobs and society shift. The relevance of certain knowledge, skills, and abilities depends on the specific social, intellectual, and physical contexts in which educational and professional actors operate. Education seeks to prepare individuals for broad activity in society. Conversely, shifts in a society's understanding of itself will affect education, its desired outcomes, and by consequence, the measurement of those outcomes. We should aim to develop tools and systems that not only conform to existing 21st century skills and technologies but we should also aim to adapt to the changes in skills and technologies we will undoubtedly see before the turn of the 22nd century.

References

American Association for the Advancement of Science. (2001). *Atlas of Science Literacy, Volume 1*. Washington, DC: American Association for the Advancement of Science.

American Association for the Advancement of Science. (2007). *Atlas of Science Literacy, Volume 2*. Washington, DC: American Association for the Advancement of Science.

Almond, R. G., DiBello, L. V., Moulder, B., & Zapata-Rivera, J. D. (2007). Modeling diagnostic assessments with Bayesian networks. *Journal of Educational Measurement, 44,* 341-359.

Almond, R. G., & Mislevy, R. J. (1999). Graphical models and computerized adaptive testing. *Applied Psychological Measurement, 23,* 223-237.

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, 5*. Retrieved from http://escholarship.bc.edu/jtla/vol1/5.

Anderson, J. (2009). *Cognitive psychology and its implications* (7th Ed.). New York, NY: Worth Publishers.

Barabasi, A. L. (2003). *Linked: How everything is connected to everything else.* New York, NY: Plume.

Baayen, H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R* . Cambridge, UK: Cambridge University Press.

Behrens, J. T., Collison, T. A., & DeMark, S. F. (2005). The seven Cs of comprehensive assessment: Lessons learned from 40 million classroom exams in the Cisco Networking Academy Program. In S. Howell and M. Hricko (Eds.), *Online assessment and measurement: Case studies in higher education, K-12 and corporate* (pp 229-245). Hershey, PA: Information Science Publishing.

Behrens, J. T., Frezzo, D. C., Mislevy, R. J., Kroopnick, M., & Wise, D. (2008). Structural, functional, and semiotic symmetries in simulation-based games and assessments. In E. Baker, J. Dickieson, W. Wulfeck, & H. F. O'Neil (Eds.), *Assessment of problem solving using simulations* (pp. 59-80). New York, NY: Earlbaum.

Behrens, J. T., Mislevy, R. J., Bauer, M., Williamson, D. M., & Levy R. (2004). Introduction to evidence centered design and lessons learned from its application in a global e-learning program. *The International Journal of Testing, 4,* 295–301.

Bell, G., & Gemmell, J. (2009). *Total recall: How the e-memory revolution will change everything.* New York, NY: Dutton.

Benkler, Y. (2007). *The wealth of networks: How social production transforms markets and freedom.* New Haven, CT: Yale University Press.

Bennett, R.E., & Bejar, I.I. (1998). Validity and automated scoring: It’s not only the scoring. *Educational Measurement: Issues and Practice, 17*(4), 9-17.

Biber, D., Conrad,S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge, UK: Cambridge University Press.

Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences.* Mahwah, NJ: Lawrence Erlbaum.

Brennan, R. L. (2001). *Generalizability theory.* New York, NY: Springer-Verlag.

Cheng, B. H., Ructtinger, L., Fujii, R., & Mislevy, R. (2010). Assessing systems thinking and complexity in science (*Large-Scale Assessment Technical Report 7*). Menlo Park, CA: SRI International. Downloaded September 11, 2010 from http://ecd.sri.com/downloads/ECD_TR7_Systems_Thinking_FL.pdf

Clark, D., Nelson, B., Sengupta, P., & D'Angelo, C. (2009, October). *Rethinking science learning through digital games and simulations: Genres, examples, and evidence.* Paper presented at a workshop of the National Academy of Science's Committee for Learning Science: Computer Games, Simulations, and Education, Washington, D.C.

Committee on Ensuring the Utility and Integrity of Research Data in a Digital Age and National Academy of Sciences. (2009). *Ensuring the integrity, accessibility, and stewardship of research data in the digital age*. Washington, DC: National Academies Press.

Corcoran, T., Mosher, F., & Rogat, A. (2009). *Learning progressions in science: An evidence based approach to reform.*. (CPRE Research Report # RR-63). New York: Consortium for Policy Research in Education & Center on Continuous Instructional Improvement, Teachers College at Columbia University .

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

De Boeck, P., & Wilson, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach.* New York, NY: Springer-Verlag.

De Nooy, W., Mrvar, A., & Batagelj, V. (2005). *Exploratory social network analysis with Pajek*. Cambridge, UK: Cambridge University Press.

DeMark, S. F. & Behrens, J. T. (2004). Using statistical natural language processing for understanding complex responses to free-response tasks. *International Journal of Testing, 4,* 371-390.

Engeström, Y. (1999) Activity theory and individual and social transformation. In Y. Engeström, R. Miettinen, & R. Punamäki (Eds), *Perspectives on Activity Theory* (pp. 19-38). Cambridge, UK: Cambridge University Press.

Freeman, L. C. (2004). *The development of social network analysis: A study in the sociology of science*. Vancouver, BC: Empirical Press.

Frezzo, D. C., Behrens, J. T., & Mislevy, R. J. (2009). Activity theory and assessment theory in the design and understanding of the Packet Tracer ecosystem. *International Journal of Learning and Media, 1*(2). doi:10.1162/ijlm.2009.0015

Frezzo, D. C., Behrens, J. T., Mislevy, R. J. (2010). Design patterns for learning and assessment: Facilitating the introduction of a complex simulation-based learning environment into a community of instructors. *Journal of Science Education and Technology, 19,* 105-114. doi:10.1007/s10956-009-9192-0

Gardner, H. E. (1987). *The mind's new science: A history of the cognitive revolution.* New York, NY: Basic Books.

Gee, J. P. (2003). *What video games have to teach us about learning and literacy*. New York, NY: Palgrave/ Macmillan.

Gries, S. (2009). *Quantitative corpus linguistics with R: A practical introduction.* London, UK: Routledge.

Gulliksen, H. (1987). *Theory of mental tests.* Hillsdale, NJ: Erlbaum.

Herrenkohl, L., & Wertsch, J. V. (1999). The use of cultural tools: Mastery and appropriation. In I. E. Sigel (Ed.), *Development of mental representation: Theories and applications* (pp. 415–435). Mahwah, NJ: Erlbaum.

Internation Data Corporation. (2008). *The diverse and exploding digital universe*. Downloaded 11/10/2010 from http://www.emc.com/collateral/analyst-reports/diverse-exploding-digital-universe.pdf

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

Kane, M., & Mitchell, R. (1996). *Implementing performance assessment: Promises, problems, and challenges.* Mahwah, NJ: Lawrence Erlbaum Associates.

Kline, R. B. (2010) *Principles and practice of structural equation modeling* (3rd ed.). New York, NY: Guilford.

Lauritzen, S., & Spiegelhalter, D. (1988). Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society B, 50*, 157-224.

Levy, F., & Murnane, R.J. (2004). *The new division of labor: How computers are creating the next job market*. Princeton, NJ: Princeton University Press.

Levy, R., & Mislevy, R. J. (2004). Specifying and refining a measurement model for a computer-based interactive assessment. *International Journal of Testing, 4,* 333-369.

Liu, M., Mislevy, R., Colker, A. M., Fried, R., & Zalles, D. (2010). A design pattern for experimental investigation *(Large-Scale Assessment Technical Report 8)*. Menlo Park, CA: SRI International. Downloaded 9/11/2010 from http://ecd.sri.com/downloads/ECD_TR8_Experimental_Invest_FL.pdf

Lord, F. M. (1980) *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

Lord, F.M., & Novick, M.R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Mager, R. (1962). *Preparing instructional objectives*. Palo Alto, CA: Fearon Publishers.

Manning, C. D., & Schuetze, H. (1999). *Foundations of statistical natural language processing.* Cambridge, MA: MIT Press.

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

Mislevy, R.J. (1994). Evidence and inference in educational assessment. *Psychometrika, 59,* 439-483.

Mislevy, R.J. (2006). Cognitive psychology and educational assessment. In R.L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 257-305). Phoenix, AZ: Greenwood.

Mislevy, R.J., Almond, R.G., & Lukas, J. (2004). A brief introduction to evidence-centered design. (CSE Technical Report 632). Los Angeles: The National Center for Research on Evaluation, Standards, Student Testing (CRESST).

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

Mislevy, R.et al.(2009). A design pattern for observational investigation assessment tasks (Large-Scale Assessment Technical Report 2). Menlo Park, CA: SRI International. Downloaded 9/11/2010 from //ecd.sri.com/downloads/ECD_TR2_DesignPattern_for_ObservationalInvestFL.pdf

Mislevy, R.J., Riconscente, M.M., & Rutstein, D.W. (2009). Design patterns for assessing model-based reasoning (*PADI-Large Systems Technical Report 6*). Menlo Park, CA: SRI International. Downloaded 9/11/2010 from http://ecd.sri.com/downloads/ECD_TR6_Model-Based_Reasoning.pdf

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2002). Design and analysis in task-based language assessment. *Language Testing, 19,* 477-496.

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

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

Nichols, P. D., Chipman, S. F., & Brennan, R. L. (Eds.). (1995). *Cognitively diagnostic assessment.* Hillsdale, NJ: Erlbaum.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference.* San Mateo, CA: Kaufmann.

Perez, C. (2003). *Technological revolutions and financial capital: The dynamics of bubbles and golden ages.* Cheltenham, UK: Edward Elgar.

Porter, M.E., & Kramer, M. R. (2002). The competitive advantage of corporate philanthropy. *Harvard Business Review, 80,* 5-16.

Reye, J. (2004). Student modeling based on belief networks. *International Journal of Artificial Intelligence in Education, 14,*1–33.

Rupp, A. A., Templin, J., & Henson, R. J. (2010). *Diagnostic measurement: Theory, methods, and applications.* New York, NY: Guilford.

Segaran, T. (2007). *Programming collective intelligence*. New York, NY: O'Reilly Media.

Schaafstal, A., and Schraagen, J.M. (2000). Training of troubleshooting: A structured, task analytical approach. In Schraagen, J. M., Chipman, S. F., and Shalin, V. L. (Eds.), *Cognitive task analysis* (pp. 57–70). Mahwah, NJ: Erlbaum.

Shaffer, D. W. (2006). Epistemic frames for epistemic games. *Computers & Education, 46,* 223-234.

Shute, V. J. (in press 2010). Stealth assessment in computer-based games to support learning. In S. Tobias, & J. D. Fletcher (Eds.), *Computer games and instruction.* 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.

Shute, V. J., Ventura, M., Bauer, M. I., & Zapata-Rivera, D. (2009). Melding the power of serious games and embedded assessment to monitor and foster learning: Flow and grow. In U. Ritterfeld, M. Cody, & P. Vorder, (Eds.), *Serious games: Mechanisms and effects* (pp. 295-321). Mahwah, NJ: Routledge, Taylor, and Francis.

Shute, V., Hansen, E., & Almond, R. (2008). You can't fatten a hog by weighing it- or can you? Evaluating an assessment for learning system called aced. *International Journal of Artificial Intelligence in Education, 18*(4), 289-316.

Spiegelhalter, D. J., Dawid, A. P., Lauritzen, S. L., & Cowell, R. G. (1993). Bayesian analysis in expert systems. *Statistical Science, 8,* 219-247.

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., Johnson, D.F., & Soller, A. (2005). Probabilities and predictions: Modeling the development of scientific problem-solving skills. *Cell Biology Education, 4,* 42–57.

Tapscott, D., & Williams, A. D. (2008). *Wikinomics: How mass collaboration changes everything.* New York, NY: Portfolio.

Tekian, A., McGuire, C. H., & McGahie, W. C. (Eds.). (1999). *Innovative simulations for assessing professional competence*. Chicago, IL: University of Illinois, Department of Medical Education.

VanLehn, K. (2008). Intelligent tutoring systems for continuous, embedded assessment. In C. A. Dwyer (Ed.), *The future of assessment: Shaping teaching and learning*, (pp. 113-138). New York, NY: Erlbaum.

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., Mislevy, R. J. Steinberg, L., & Thissen, D. (2000). *Computerized adaptive testing: A primer*. New York, NY: Routledge.

Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications.* Cambridge, UK: Cambridge University Press.

West, P., Rutstein, D.W., Mislevy, R..J., Liu, J., Levy, R., DiCerbo, K.E., Crawford, A., Choi, Y., Chappel, K., & Behrens, J.T. (2010). *A Bayesian network approach to modeling learning progressions.* (CRESST Research Report 776). Los Angeles: The National Center for Research on Evaluation, Standards, Student Testing (CRESST), Center for Studies in Education, UCLA.

Wiley, D. E., & Haertel, E. H. (1996). Extended assessment tasks: Purposes, definitions, scoring, and accuracy. In M. B. Kane & R. Mitchell (Eds.), *Implementing performance assessment: Promises and challenges* (pp. 61-89). Mahwah, NJ: Lawrence Erlbaum Associates.

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,* 333–369.

Wise-Rutstein, D. (2005, April). *Design patterns for assessing troubleshooting in computer networks.* Presentation at the annual meeting of the American Education Research Association, San Francisco, CA.

Wolfram, S. (2002). *A new kind of science* (1st ed.). Champaign, IL: Wolfram Media.

Wright, B. D., & Stone, M. H. (1979). *Best test design.* Chicago, IL: MESA Press.