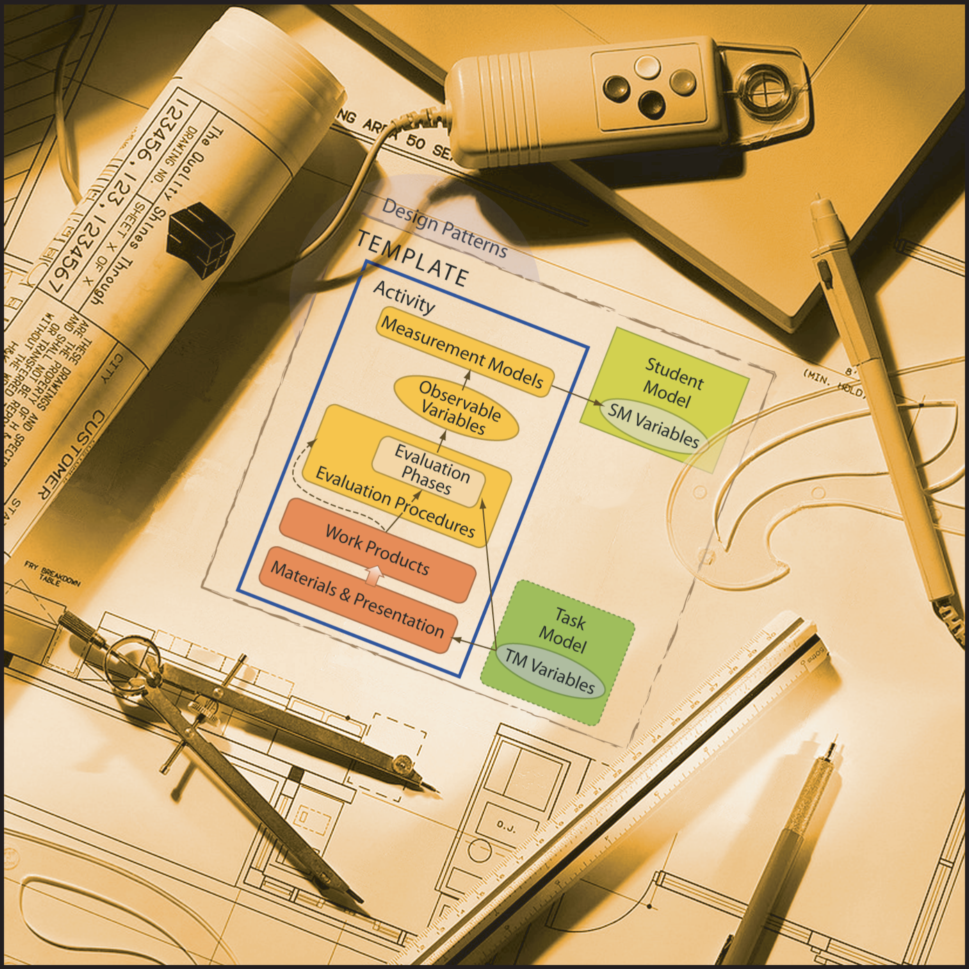


PADI Technical Report 2 | September 2003

Leverage Points for Improving Educational Assessment

PADI | Principled Assessment Designs for Inquiry

Robert J. Mislevy, University of Maryland
Linda S. Steinberg, University of Pennsylvania
Russell G. Almond, Educational Testing Service
Geneva D. Haertel, SRI International
William R. Penuel, SRI International

Report Series Published by SRI International

SRI International
Center for Technology in Learning
333 Ravenswood Avenue
Menlo Park, CA 94025-3493
650.859.2000
http://padi.sri.com

PADI Technical Report Series Editors
Alexis Mitman Colker, Ph.D. *Project Consultant*
Geneva D. Haertel, Ph.D. *Co-Principal Investigator*
Robert Mislevy, Ph.D. *Co-Principal Investigator*
Klaus Krause. *Technical Writer/Editor*
Lynne Peck Theis. *Documentation Designer*

PRINCIPLED ASSESSMENT DESIGNS FOR INQUIRY
TECHNICAL REPORT 2

Leverage Points for Improving Educational Assessment

Prepared by:
Robert J. Mislevy, University of Maryland
Linda S. Steinberg, University of Pennsylvania
Russell G. Almond, Educational Testing Service
Geneva D. Haertel, SRI International
William R. Penuel, SRI International

Acknowledgment
PADI is supported by the Interagency Educational Research Initiative (IERI) under grants REC-0089122 (PADI Planning Grant) and REC-0129331 (PADI Implementation Grant).

Disclaimer
Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

CONTENTS

FIGURES

ABSTRACT

Advances in cognitive psychology deepen our understanding of how students gain and use knowledge. Advances in technology make it possible to capture more complex performances in assessment settings, by including, for example, simulation, interactivity, collaboration, and constructed response. The challenge is in knowing just how to put this new knowledge to work. Familiar schemas for designing and analyzing tests produce assessments that are useful because they are coherent, within the constraints under which they evolved. Breaking beyond the constraints requires not only the means for doing so (through the advances mentioned above) but schemas for producing assessments that are again coherent; that is, assessments that may indeed gather complex data to ground inferences about complex student models, to gauge complex learning or evaluate complex programs--but which build on a sound chain of reasoning from what we observe to what we infer. This presentation first reviews an evidence-centered framework for designing and analyzing assessments. It then uses this framework to discuss and to illustrate how advances in technology and in education and psychology can be harnessed to improve educational assessment.

Introduction

Interest in complex and innovative assessment is expanding for several reasons. Advances in cognitive and educational psychology broaden the range of things we want to know about students, and possibilities for observable variables to give us evidence (Glaser, Lesgold, & Lajoie, 1987). We have opportunities to put new technologies to use in assessment, to create new kinds of tasks, to bring them to life, and to interact with examinees (Bennett, 1999; Quellmalz & Haertel, 1999). We are called on to investigate the success of technologies in instruction, even as they target knowledge and skills that are not well measured by conventional assessments. But how do we design complex assessments so they provide the information we need to achieve their intended purpose? How do we make sense of the complex data they may generate?

This presentation is based on two premises. First, the principles of evidentiary reasoning that underlie familiar assessments are a special case of more general principles. Second, these principles can help us design and analyze new kinds of assessments, with new kinds of data, to serve new purposes.

The first half of this report reviews an “evidence-centered” (Schum, 1994) framework for designing assessments (Mislevy, Steinberg, & Almond, 2002). The second half discusses, through the lens of this framework, how and where advances in cognitive psychology and technology can be brought to bear to improve assessment. We draw on three examples to illustrate ideas throughout the report: (1) a familiar standardized test, the GRE; (2) a prototype simulation-based assessment of problem solving in dental hygiene for the Dental Interactive Simulations Corporation (DISC) (Mislevy, Steinberg, Breyer, Almond, & Johnson, 1999, 2002); and (3) an online performance task, the MashpeeQuest, designed to evaluate students’ information analysis skills, as part of Classroom Connect’s AmericaQuest instructional program (Penuel & Shear, 2000).

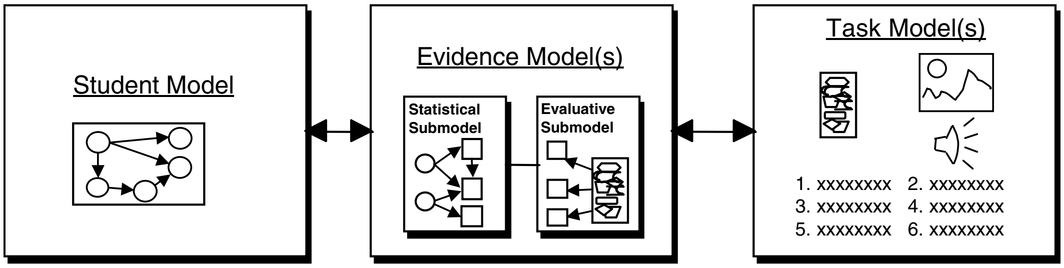
Evidence-Centered Assessment Design

There are two kinds of building blocks for educational assessment. Substantive building blocks concern the nature of knowledge in the domain of interest, how students learn it, and how they use their knowledge. Evidentiary-reasoning building blocks concern what and how much we learn about students' knowledge from what they say and do. How do we assemble these building blocks into an assessment? This section reviews Mislevy, Steinberg, and Almond's (2002) "conceptual assessment framework" (CAF). In the following section, "Leverage Points for Improving Assessment," we use the structure of the CAF to discuss where and how advances in psychology and technology can be put to work to improve the practice of assessment.

Figure 1 is a high-level schematic of the CAF, showing three basic models we suggest must be present, and must be coordinated, to achieve a coherent assessment. A quote from Messick (1994, p. 16) serves to introduce them:

> A construct-centered approach [to assessment design] 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.

Figure 1. Three Basic Models of Assessment Design



The Student Model

"What complex of knowledge, skills, or other attributes should be assessed?" Configurations of values of student-model variables are meant to approximate, from some perspective about skill and knowledge in the domain, certain aspects of possible configurations of skill and knowledge real students have. The perspective could be that of behaviorist, trait, cognitive, or situative psychology. But whatever the perspective, we encounter the evidentiary problem of reasoning from limited evidence. Student-model variables are the terms in which we want to talk about students, to determine evaluations, make decisions, or plan instruction—but we don't get to see the values directly. We just see what the students say or do and must use this as evidence about the student-model variables.

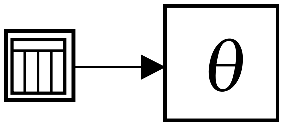
The student model in Figure 1 depicts student-model variables as circles. The arrows connecting them represent important empirical or theoretical associations. These variables and associations are implicit in informal applications of reasoning in assessment, such as a one-to-one discussion between a student and a tutor. In the more formal applications discussed in this report, we use a probability model to manage our knowledge about a given student's (inherently unobservable) values for these variables at any given point in time. We express our knowledge as a probability distribution, which can be updated in light of new evidence. In particular, the student model takes the form of a fragment of a Bayesian inference network, or Bayes net (Jensen, 1996).

An understanding of competence in the domain is necessary for determining the number and nature of the student-model variables to use in a given application. This determination of number will also depend on the purpose of the assessment. A single variable that characterizes overall proficiency in a particular domain might suffice in an assessment meant to support only a summary pass/fail decision. But a coached practice system to help students develop that same proficiency would require a finer-grained student model, in order to monitor particular aspects of skill and knowledge for which feedback is available. When the purpose is program evaluation, the grain size and the nature of the student-model variables should reflect ways in which a program may enjoy more or less success or promote students' learning in some ways as opposed to others. The purpose of the example in the MashpeeQuest assessment is to gather information about students' information-gathering and synthesis skills in a technological environment. It follows that the student model should include variables that concern aspects of these skills, and these variables will be defined more concretely by the kinds of observations we will posit as constituting evidence about them.

It requires further thought to decide whether to include student-model variables for aspects of these skills as they are used in nontechnological situations, as they are evidenced by observations from nontechnological situations. There are two reasons one might include student-model values thusly conceived and revealed. Both revolve around the intended purpose of an assessment. First, if we want talk about differential impacts in different environments, we must be able to distinguish skills as they are used in different technological environments. This might be done with a multivariate student model with variables that disentangle such effects from the same complex performances, or with multiple but distinct assessments with different sources of evidence and each with its own student-model variables. Second, if we want to compare students in the targeted instructional program with students not in that program, we will not be able to obtain evidence from the latter with ways of collecting evidence that depend on being familiar with technologies specific to the program.

Example 1: The GRE. Figure 2 depicts the student model that underlies most familiar assessments. A single variable, typically denoted θ, represents proficiency in a specified domain of tasks. We use as examples the paper and pencil (P&P) and the computer-adaptive test (CAT) versions of the Graduate Record Examination (GRE), which comprise domains of items for Verbal, Quantitative, and Analytic Writing skills. The small table in the square in front of this student-model (SM) variable represents the probability distribution that expresses current belief about a student's unobservable status. At the beginning of an examinee's assessment, the probability distribution representing a new student's status will be uninformative. We update it in accordance with responses to GRE Verbal test items.

Figure 2. The Student Model for a GRE Verbal Measure

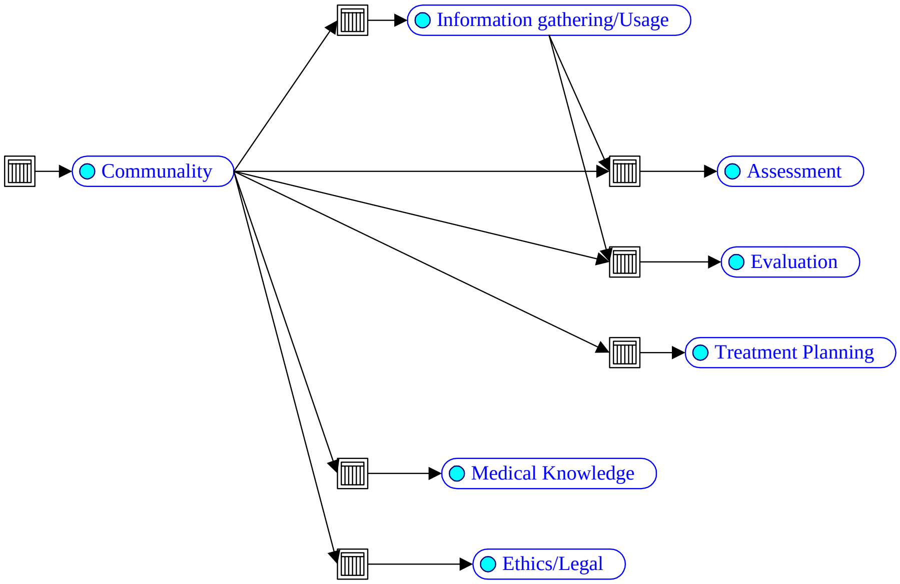


We can describe this model in terms of Bayes nets. In assessment, a Bayes net contains both student-model variables, the inherently unobservable aspects of knowledge or skill about which we want to draw inferences, and observable variables, for which we can ascertain values directly, and which are dependant on the probability of the student-model variables. The student-model variables are a fragment of this complete network. Another kind of fragment contains one or more observable variables and pointers to the student-model variables they depend on. As discussed in the section on evidence models, we can combine ("dock") the student-model (SM) Bayes net fragment with an appropriate evidence-model (EM) fragment when we want to update our beliefs about the student-model variables in light of data (Almond & Mislevy, 1999).

Example 2: DISC. Educational Testing Service (ETS) is working with The Chauncey Group International (CGI) to develop a scoring engine for a prototype of a simulation-based assessment of problem solving in dental hygiene, under contract with the Dental Interactive Simulations Corporation (DISC). They are working through student, evidence, and task models with DISC, and consequently examining the implications for the simulator. Two considerations shaped the student model for the prototype assessment. The first was the nature of skills DISC wanted to focus on: the problem-solving and decision-making skills a hygienist employs on the job. The second was the purpose of the assessment: a licensure decision, with some supplementary information about strengths and weaknesses. This paper therefore refers to the student model described below as an "overall proficiency + supplementary feedback" student model.

Adapting cognitive task analysis methods from the expertise literature (Ericsson & Smith, 1991), the ETS, CGI, and DISC researchers captured and analyzed protocols from hygienists at different levels of expertise as they solved a range of tasks in the domain (Johnson et al., 1998). General characterizations of patterns of behavior were abstracted—a language that could describe solutions across subjects and cases, not only in the data at hand but in the domain of dental hygiene decision-making problems more broadly. An example was "Using disparate sources of information." Novice hygienists were usually able to note important cues on particular forms of information, such as shadows on radiographs and bifurcations on probing charts, but they often failed to generate hypotheses that required integrating cues across different forms. Student-model variables were defined that would characterize a hygienist's tendency to demonstrate these indicators, both overall and broken down into a smaller number of facets. Scores on facets could also be reported to students. Figure 3 is a simplified version of the student model currently in use.

Figure 3. Simplified DISC Student Model



The ovals in Figure 3 are the SM variables. The two toward the upper right are Assessment, or proficiency in assessing the status of a new patient, and Information-gathering/Usage. Information-gathering/Usage is further elaborated into variables for knowing how and where to obtain information, being able to generate hypotheses that would guide searches and interpretations, and knowing how to gather information. The information about these variables would help confirm or refute hypotheses.

Example 3: MashpeeQuest. Our third example was an online performance task designed by researchers from SRI International to evaluate Classroom Connect's AmericaQuest instructional program. AmericaQuest aims to help students learn to develop persuasive arguments, supported by evidence they acquire from the course's Web site or from their own research. MashpeeQuest poses a problem that gives students an opportunity to put these skills to use in a Web-based environment that structures their work.

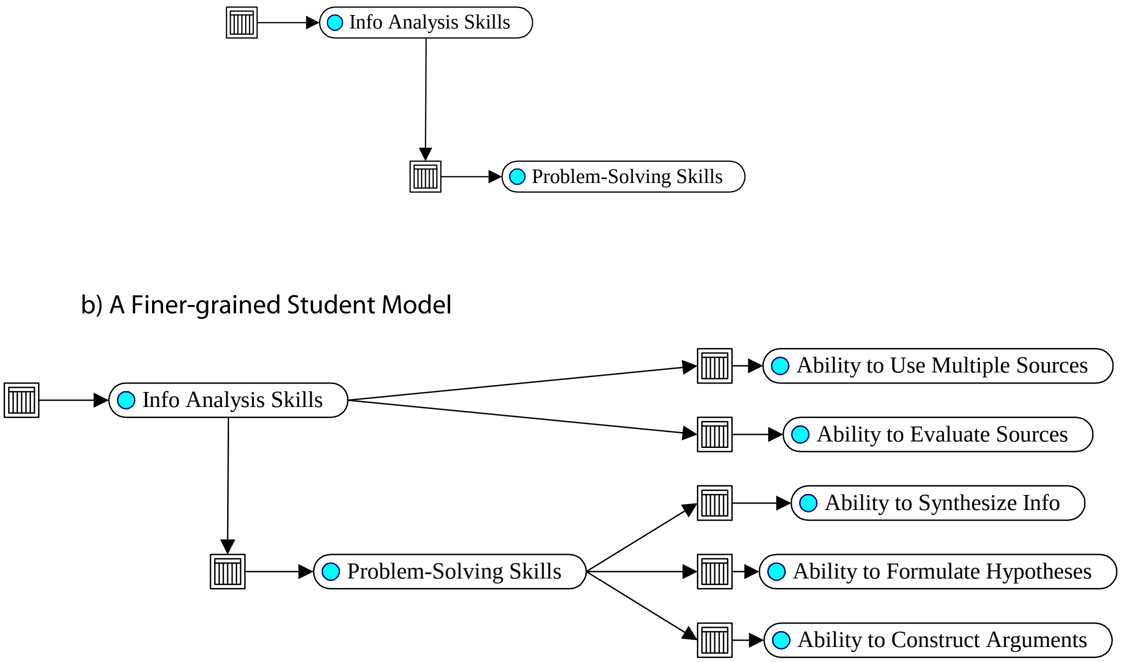
The design of the MashpeeQuest performance task was motivated by the goals of the evaluation. It assesses a subset of the skills that the AmericaQuest program is meant to foster:

- *Information analysis skills.* Ability to analyze and synthesize information from a variety of sources; ability to evaluate/critique both content and sources.

- *Problem-solving skills.* Ability to synthesize disparate ideas through reasoning in a problem-solving context; ability to offer reasoned arguments rather than brief guesses; ability to formulate creative, well-founded theories for unsolved questions in science and history.

Figure 4 illustrates two possible student models that are consistent with the preceding description. They differ in their specificity, or grain size. The first contains only two variables and would be used to accumulate information about students in terms of just information analysis skills and problem-solving skills. The arrow between them indicates that they may be correlated in the population of students being addressed. The second student model includes variables for subskills, so that evidence can be accumulated separately for them and used to identify for students or teachers more specific areas of strength or difficulty. Deciding which of the two models to use would require (1) weighing the more detailed information in the finer-grained model against its lower accuracy, and (2) examining the empirical correlation among the subskills, since the more highly they are correlated the less is gained by modeling them explicitly.

Figure 4. Two Possible MashpeeQuest Student Models


a) A Coarse-grained Student Model

b) A Finer-grained Student Model
Info Analysis Skills
Problem-Solving Skills
Info Analysis Skills
Ability to Use Multiple Sources
Ability to Evaluate Sources
Ability to Synthesize Info
Problem-Solving Skills
Ability to Formulate Hypotheses
Ability to Construct Arguments

The effective meaning of any of these student-model variables will be determined by choices about the observations that are deemed to constitute evidence about them. In the MashpeeQuest task, students will have to weigh evidence they might find in on-line visits to cities in the northeastern United States to help decide a court case involving recognition for the Mashpee Wampanoags, a Native American tribe in Massachusetts. A band of people claiming Wampanoag ancestry have been trying for more than 20 years to gain recognition from the federal government as a tribe that still exists. In 1978, a federal court ruled against the Mashpee Wampanoags' claim, arguing that the tribe could not prove that it had a continuous stake on territory in Mashpee. The tribe is seeking recognition a second time in court. The assessment asks students to take a position on the case and to identify places where a Quest expedition team should go based on information about the kinds of evidence they might find there. Students are asked to investigate the evidence, select sites that provide evidence to support their claim, and justify their choices on the basis of the evidence. In addition, they are asked to identify one place to go to find evidence that doesn’t support their claim, and to address how their theory of what happened to the Mashpee Wampanoags is still justified.

The developers of the Mashpee task had to tackle the issue of how to define student-model variables in the evaluation of a technology-based program. This task was designed specifically for use with students who have become familiar with the vocabulary and affordances of the technological environment of AmericaQuest. It obtains evidence about how well they can apply the skills they presumably have been developing in the AmericaQuest environment, as well as on other problems. This task’s role in the evaluation is to provide evidence about whether the students in the program can in fact use the skills they have been working on, rather than to compare these students with other students from different programs, or even with themselves before they began the program. Other components of the evaluation have been designed to produce evidence that can be compared across groups whether or not they are familiar with the environment and conventions of AmericaQuest.

The Evidence Model

“What behaviors or performances should reveal those constructs,” and what is the connection? The evidence model lays out our argument about why and how the observations in a given task situation constitute evidence about student-model variables. Figure 1 shows two parts to the evidence model, the *evaluative submodel* and the *statistical submodel*. The statistical submodel updates the student model in accordance with the values of these features, thus synthesizing the evidentiary value of performances over tasks (Mislevy & Gitomer, 1996). The evaluative submodel extracts the salient features of the work product: whatever the student says, does, or creates in the task situation.

The Evaluative Submodel. In the icon for the evaluative submodel in Figure 1, the work product is a rectangle containing a jumble of complicated figures at the far right. It is a unique human production, as simple as a mark on an answer sheet or as complex as a series of evaluations and treatments in a patient-management problem. The squares coming out of the work product represent “observable variables,” the evaluative summaries of the key aspects of the performance. The evidence rules map unique human products into a common interpretive framework. These mappings can be as simple as determining whether the mark on an answer sheet is the correct answer or as complex as an expert’s evaluation of multiple aspects of a patient-management solution. They can be automatic or require human judgment.

An evidence rule taken from the GRE example follows:

```
IF the response selected by the examinee matches the response marked as the
key in the database,

THEN the item response IS correct

ELSE the item response IS NOT correct.
```

What is important about this evidence rule is that a machine can carry it out. This technological breakthrough slashed the costs of testing in the 1940s. But the new technology did not change the essential nature of the evidence or the inference. It was used to streamline the process by modifying the student's work product to a machine-readable answer sheet and having a machine rather than a human apply the evidence rules.

In the second example, based on DISC, the cognitive task analysis (CTA) produced "performance features" that characterize patterns of behavior and differentiate levels of expertise. These patterns were the basis of generally defined, reusable observed variables. The evidence models themselves are assemblies of student-model variables and observable variables, including methods for determining the values of the observable variables and updating student-model variables accordingly. A particular assessment case will use the structures of one or more evidence models, fleshed out in accordance with specifics of that case.

The evaluative submodel of an evidence model concerns the mappings from examinees' unique problem solutions into a common framework of evaluation—that is, from work products to values of observable variables. The constant elements in the evaluative submodels for tasks that are built to conform to the same evidence model are the identification and formal definition of observable variables, and generally stated "proto-rules" for evaluating their values. Adequacy of examination procedure is an aspect of any assessment of any new patient. For example, we can define a generally stated evaluative framework to describe how well an examinee has adapted to whatever situation is presented. The elements that are customized to particular cases are case-specific rules, or rubrics, for evaluating values of observables—instantiations of the proto-rules tailored to the specifics of each case. The unique features of a particular virtual patient's initial presentation in a given assessment situation determine what an examinee ought to do in the assessment, and why.

The MashpeeQuest task, which is the third example, requires students to demonstrate a particular set of information analysis and problem-solving skills. These skills comprise the student model. While the task provides only a single problem context in which students can demonstrate these skills, it provides multiple opportunities for students to demonstrate different aspects of information analysis skill, in different ways, in different parts of the problem. The observable variables, identified in using MashpeeQuest and defined to be evidence of information analysis skills, all demonstrate more generally cast skills one needs to use the Internet to conduct research or inquiry: comparing information from multiple sources by browsing and reading different Web links, constructing texts that compare information gleaned from these sources, and evaluating the credibility of that information. The observables evidencing problem-solving skills are specific to the AmericaQuest instructional program, but all have strong parallels to the argumentation skills required of students in other innovative Web-based learning programs (e.g., Linn, Bell, & Hsi, 1999). These include using information on the Internet as clues to solve a discrete problem, and generating theories, based on consideration of evidence and counterevidence, related to a controversy in history and anthropology.

Technology plays two roles in the evaluative component of the evidence model for the MashpeeQuest task. The first is conceptual: the information analysis skills to be assessed and the behaviors that evidence them are embedded within the Web-based assessment environment. The MashpeeQuest task intentionally takes a specific context for analyzing information—the World Wide Web—and tests a model of information analysis that involves performances specific to using the Web for research and inquiry (e.g., clicking through different links, inferring the validity of sources from specific aspects of the Web page). The second is more operational: because actions take place in a technological environment, some of the observables can be evaluated automatically. Evidence rules for the observables "Number of sources" and "Time per source" are as straightforward as those for the GRE paper and pencil test. Other observables are better evaluated by people. For example, student performance on subtasks requiring information analysis would be scored by human raters using a rubric that evaluates students' "Discussion of Coherence" and "Discussion of Credibility" of the sites they visited.

The Statistical Submodel. In the icon for the statistical submodel in Figure 1, the observables are modeled as depending on some subset of the student-model variables. Item response theory, latent class models, and factor analysis are examples of psychometric modeling approaching in which values of observed variables depend in probability on unobservable variables. These can be expressed as special cases of Bayes nets and extend the ideas as appropriate to the nature of the student-model and observable variables (Almond & Mislevy, 1999; Mislevy, 1994). In complex situations, statistical models from psychometrics can play crucial roles as building blocks. These models evolved to address certain recurring issues in reasoning about what students know and can do, given what we see them do in a limited number of circumscribed situations, often captured as judgments of different people who may not agree in their evaluations.

Figure 5 shows the statistical submodel of the evidence model used in the GRE CAT, an item response theory (IRT) model. On the left is a fragment of a Bayesian inference network for updating the probability distribution of the student's proficiency parameter given a response to a particular item j. The distribution between the student-model proficiency variable θ and the item response X_j represents the conditional probability distribution of X_j given θ. When it is time to make an observation, this fragment is "docked" with the SM Bayes net fragment to form a complete probability model for θ and X_j jointly. On the right of the figure is a library of all items that could be given, along with the structures necessary to dock any one with the student model in order to incorporate the evidence its response contributes. Further discussion of the statistical aspects of this process can be found in Mislevy, Almond, Yan, & Steinberg (1999).

Figure 5. The Statistical Submodel of the Evidence Model in the GRE CAT

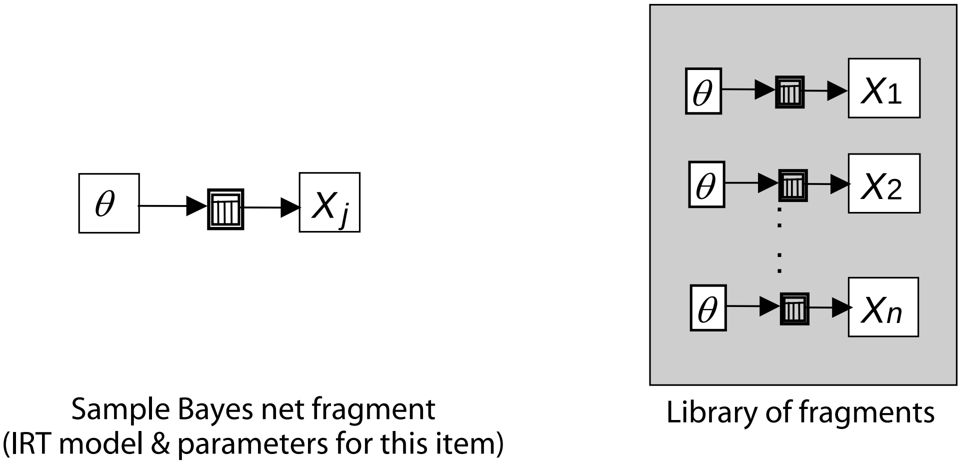
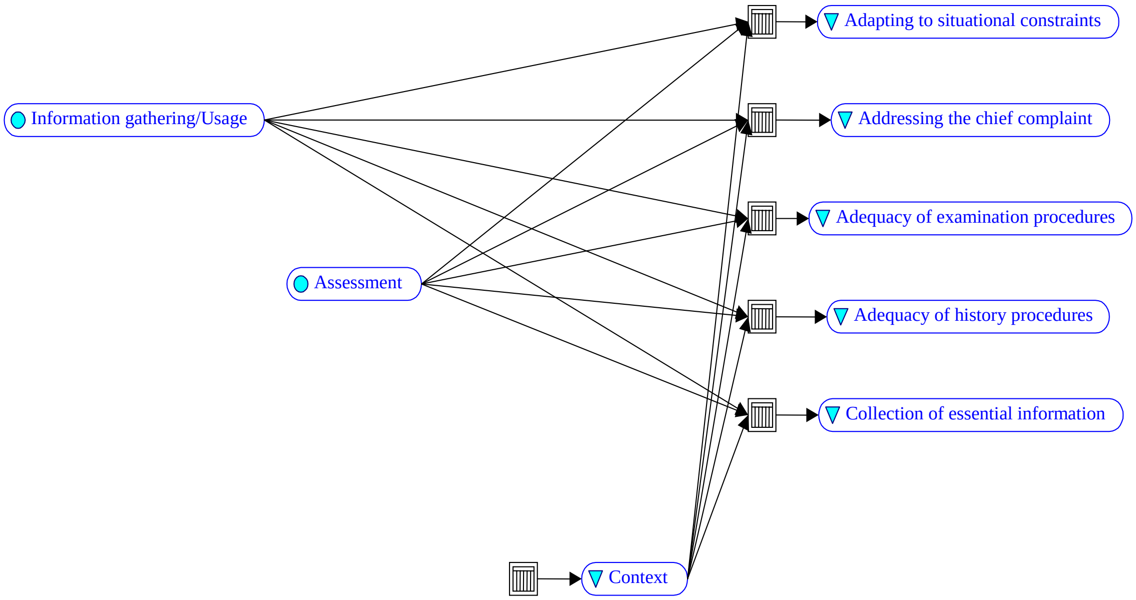

θ
X j
Sample Bayes net fragment
(IRT model & parameters for this item)
θ
X1
θ
X2
θ
Xn
Library of fragments

Figure 6 shows the Bayes net fragment that comprises the statistical submodel of one particular evidence model taken from the DISC example. It concerns gathering patient information when assessing a new patient's status. At the far left are student-model variables we posit as driving performance in these situations: *Assessment* of new patients and *Information-gathering/Usage*. The nodes on the right are generally-defined observable variables. Two of them are *adapting to situational constraints* and *adequacy of examination procedures* (in terms of how well their rationale is grounded). In a specific case, the values of the observable variables are the result of applying rubrics that have been tailored from the general rubric, for this observable, to the particulars of the case. Figure 7 shows how this evidence model Bayes net fragment is "docked" with the student-model fragment when an examinee is working in a situation that has been constructed to conform with this evidence model.

Figure 6. The Bayes Net Fragment in an Evidence Model in DISC


Adapting to situational constraints
Information gathering/Usage
Addressing the chief complaint
Adequacy of examination procedures
Assessment
Adequacy of history procedures
Collection of essential information
Context

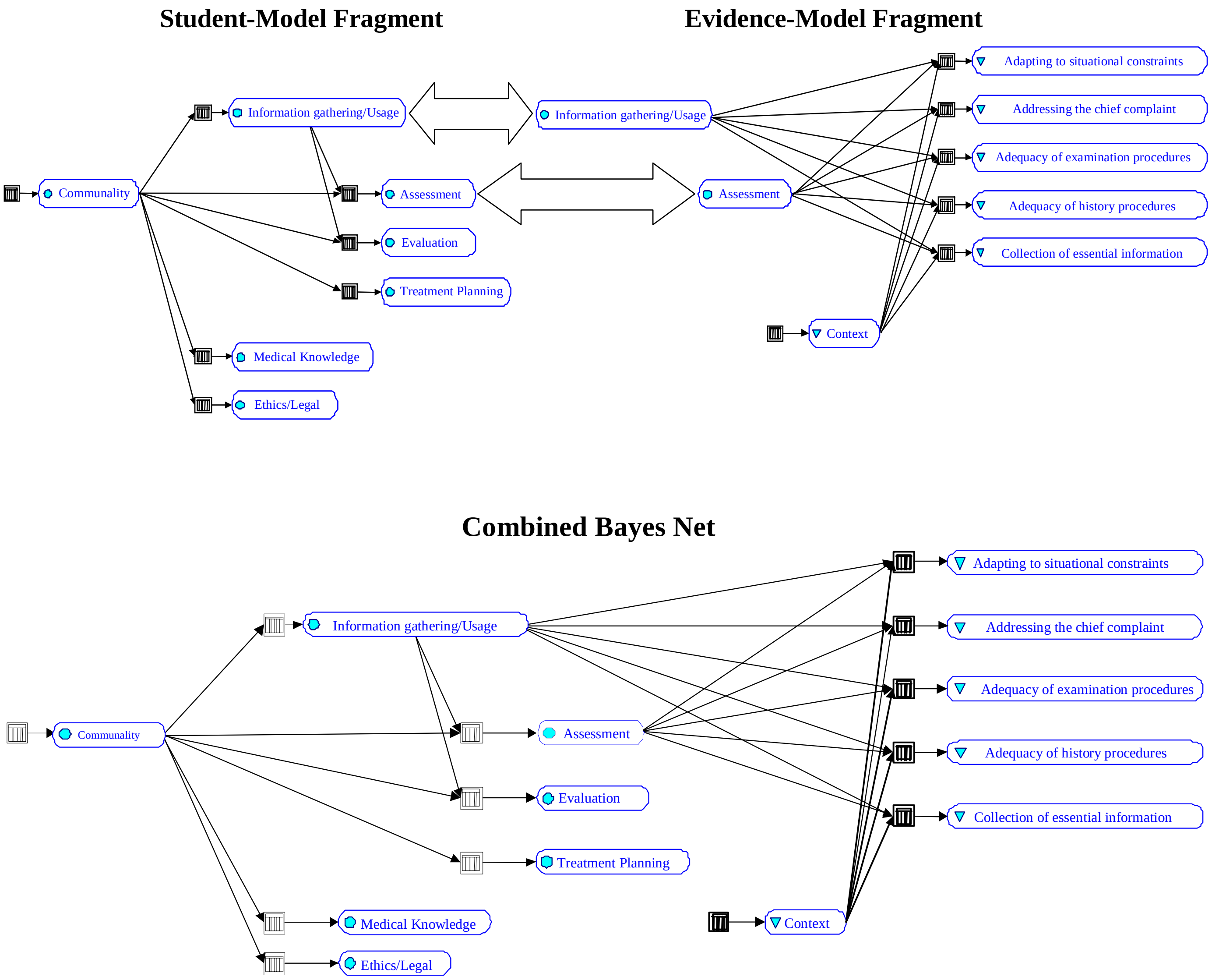
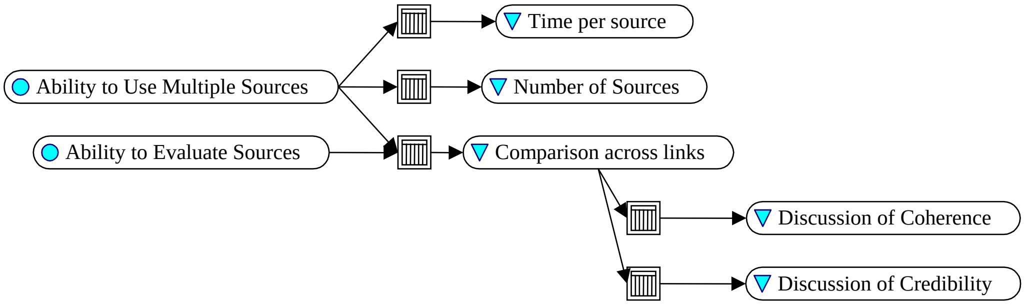

Student-Model Fragment
Evidence-Model Fragment
Information gathering/Usage
Communality
Assessment
Evaluation
Treatment Planning
Medical Knowledge
Ethics/Legal
Information gathering/Usage
Assessment
Context
Adapting to situational constraints
Addressing the chief complaint
Adequacy of examination procedures
Adequacy of history procedures
Collection of essential information
Combined Bayes Net
Information gathering/Usage
Communality
Assessment
Evaluation
Treatment Planning
Medical Knowledge
Ethics/Legal
Context
Adapting to situational constraints
Addressing the chief complaint
Adequacy of examination procedures
Adequacy of history procedures
Collection of essential information

Figure 7. Docking Student-Model and Evidence-Model Bayes Net Fragments in DISC

Figure 8 depicts the statistical submodel of the evidence model related to student information analysis skills assessed in a hypothetical family of tasks like the MashpeeQuest task. The focus is on measuring student performance in the context of a problem that requires students to read, interpret, and use information on the Web to solve a problem like those presented in the AmericaQuest program. At the left of the figure are two variables from the finer-grained student model introduced above, *Ability to use multiple sources* and *Ability to evaluate sources* that are parts of *Information gathering/usage*. These parent variables drive the probabilities of the observable variables in the middle of the figure and the lower right. We see that *Ability to use multiple sources* is informed by the observable variables *Time per source*, *Number of sources*, and [quality of] *Comparison across links*. *Number of sources* could have as many values as there are links in the task. Because no prior information is given to students about what sources are more likely to have useful information, more sources considered is taken as evidence of better information analysis skills. *Time per source* could have any number of values from a few seconds to several minutes. Here, one would see if students were simply "clicking through" without reading a particular link. The time spent is an important counter balance to the number of sources considered, since time spent is an (imperfect) indicator of whether students actually read the text on the links they used. [Quality of] *Comparison across links* is actually a composite of two ratings of the same student responses, namely, evaluations of how well they discussed the *Coherence* and the *Credibility* of the sites they visited—key features of effective information analysis, according to experts in this domain (Wineburg, 1998).

Figure 8. A MashpeeQuest Evidence Model


Time per source
Ability to Use Multiple Sources
Number of Sources
Ability to Evaluate Sources
Comparison across links
Discussion of Coherence
Discussion of Credibility

We also see that the student-model variable *Ability to evaluate sources* is informed by *Comparison across links*. *Ability to evaluate sources* is not modeled as informed by *Number of sources* or *Time per source*, although students' inability to access sites would surely prevent them from providing evaluations. For this reason, the structure of the conditional probability distribution for this observable would indicate that at least some ability to gather information across sites would be required, in addition to evaluative skill, to have a high probability of good ratings on this observable. One could in principle get evidence about *Ability to evaluate sources* unconfounded by students' ability to find them and analyze the information they contained by presenting subtasks in which students were simply presented sites and synopses of them and asked to evaluate their coherence and credibility.

The Task Model

"What tasks or situations should elicit those behaviors?" A task model provides a framework for constructing and describing the situations in which examinees act. The model includes specifications for the environment in which the student will say, do, or produce something—for example, characteristics of stimulus material, instructions, help, tools, affordances. It also includes specifications for the work product, the form in which what the student says, does, or produces will be captured. Assigning specific values to task-model variables, and providing materials that suit the specifications there given, produces a particular task. A task thus describes particular circumstances meant to provide the examinee an opportunity to act in ways that produce information about what they know or can do more generally. Distinct and possibly quite different, evidence rules could be applied to the same work product from a given task. Distinct and possibly quite different student models, befitting different purposes or derived from different conceptualizations of proficiency, could be informed by data from the same task.

A task model in the GRE describes a class of test items. There is a correspondence between task models and GRE "item types" (e.g., sentence completion, passage comprehension, quantitative comparison). These item types require different task models, because different sets of variables are needed to describe their distinct kinds of stimulus materials and presentation formats, and different features may be important in modeling item parameters or controlling item selection.

Task-model (TM) variables for the DISC prototype specify information the simulator needs for the virtual patient and features that will evoke particular aspects of skill and knowledge. A test developer can create a case by first referring to a matrix that cross-references student-model variables, evidence models that can be used to get information about the patients, and task models around which tasks can be constructed to provide that evidence. Once a task model is selected, it is fleshed out with particulars to create a new virtual patient.

Task-model variables that describe the patient include, as examples, Age, Last Visit, Reason for Last Visit, Symptoms of Abuse/Neglect, Demeanor, and Risk for Medical Emergency. Some of these are important to focus on aspects of proficiency the CTA revealed. Risk for Medical Emergency, for example, should be set to "low" or "none" for cases in which evidence about Medical Knowledge is not sought, but values of "moderate" or "high" necessitate the use of evidence models that do include Medical Knowledge as student-model parents.

Task models also include specifications for work products. The simulator records the sequence of actions an examinee takes, which can then be parsed by evidence rules. Several of the performance features that emerged from the CTA concerned intermediate mental products, such as identification of cues, generation of hypotheses, and selection of tests to explore conjectures—steps that are usually not manifest in practice, but that directly involve central knowledge and skills for problem solving in dental hygiene. Work products that require the examinee to make such steps explicit will capture more direct evidence of the thinking behind a solution than the sequence of actions will. Following patient assessment, for example, the examinee will fill out a summary form that requires synthesized findings in a form similar to commonly used insurance forms.

In the assessment designed for AmericaQuest, a number of features of the task model are not content specific, but they are subject-matter specific. These kinds of problems should involve the consideration of historical and archaeological evidence, as AmericaQuest does. Such consideration does not necessitate a focus on the Mashpee Wampanoag or any other Native American tribe per se. The problem statement should ask students to formulate a hypothesis and back it with evidence gathered from information available to them in the Web-based assessment environment, as they do in AmericaQuest and as specified in the student model of problem-solving skill. The task model would vary if students were asked to display analysis skills in ways other than stating a hypothesis and supplying Web-based evidence, for then both the kinds of material made available to students and the kinds of work products they produced could differ.

There are important ways one could vary the task to isolate particular skills identified in the student model. At present, the different links on MashpeeQuest do not all contain evidence in support of one hypothesis or another about the Mashpee Wampanoag. Some links contain evidence suggesting the tribe disappeared, while others contain evidence suggesting the tribe has maintained its traditions and culture despite generations of acculturation to American ways of life. If one were interested solely in comparison of multiple sources of information—and not in whether students could formulate ideas about the coherence of ideas across links or sources—one could vary the particular links so that students were simply accumulating different pieces of evidence in support of one particular hypothesis. All the links, for example, could support the idea that the Mashpee Wampanoag were in fact a tribe with a continuous historical existence, and the task for students would be to draw evidence to support that theory from as many different sources or links as possible. The task model could thus be defined to include variables about the number of sources available, the degree of ambiguity among them, and the variation in quality and credibility of the sources. By varying these features systematically in different contexts, the assessment designer could produce a family of Web-based investigations that varied in predictable ways as to difficulty and the skills they emphasized.

Leverage Points for Improving Assessment

This has been a quick tour of a schema for the evidentiary-reasoning foundation of assessments. It gives us some language and concepts for talking about this central core of assessment, not only for familiar forms and uses of assessment but for new forms and uses. We can use this framework to discuss ways we can take advantage of advances in psychology and technology.

Leverage Points for Psychology

While the familiar practices of assessment and test theory originated under the regimes of trait and behaviorist psychology, contemporary views of learning and cognition fit more comfortably into the headings of cognitive and situative psychology (Greeno, Collins, & Resnick, 1996). The cognitive perspective includes both the constructivist tradition originated by Piaget and the information-processing tradition developed by Newell and Simon (1972), Chomsky, and others. The focus is on patterns and procedures individuals use to acquire knowledge and put it to work. The situative perspective focuses on the ways individuals interact with other people in social and technological systems, so that learning includes becoming attuned to the constraints and affordances of these systems (e.g., Rogoff, 1984). In this report, we use the term *cognitive psychology* broadly to encompass both of these perspectives.

As Messick pointed out, in designing an assessment we start with the questions of what we want to make inferences about and what we need to see to support those inferences. From the perspective of trait psychology (the approach that produced the GRE), the targets of inference were traits that presumably influenced performance over a wide range of circumstances, and samples of those circumstances were needed—the cheaper the better, since the specifics of domains and tools were noise rather than signal. From the perspective of cognitive psychology (which generated our other two examples), the targets of inference are cast in terms of the patterns, skills, and knowledge structures that characterize developing expertise. This perspective shapes design decisions at several points in the three models that comprise the conceptual assessment framework (CAF).

The character and substance of the student model. How we conceive of students' knowledge and how it is acquired helps us frame our targets of inference—that is, the ways in which we will characterize what students know and can do. Glaser, who has long advocated the value of a cognitive perspective in assessment, makes the following case:

> At various stages of learning, there exist different integrations of knowledge, different degrees of procedural skill, differences in rapid access to memory and in representations of the tasks one is to perform. The fundamental character, then, of achievement measurement is based upon the assessment of growing knowledge structures, and related cognitive processes and procedural skills that develop as a domain of proficiency is acquired. These different levels signal advancing expertise or passable blockages in the course of learning. (Glaser, Lesgold, & Lajoie, 1987, p. 77)

The DISC project provides a first example of how we can characterize what students can know and do. The CTA provided insights into the kinds of knowledge hygienists used, and thus into the dimensions along which we might wish to characterize their levels and degrees of proficiency. Recall, though, that this information is necessary, but not sufficient, for defining the variables in a student model. Equally important is the purpose the assessment is intended to serve. If DISC only wanted to make a single pass/fail decision on an overall index of proficiency, a student model with a single variable might still be used to characterize an examinee. The DISC assessment designers might even use the same task models that we outlined above for our “overall decision + supplementary feedback” purposes. If DISC wanted to build an intelligent tutoring system, they might need a far more detailed student model, again consistent with the same conception of expertise but now detailed enough to capture and manage belief about many aspects that are more fine-grained, of knowledge and skills. Only at the fine-grained level would DISC assessment designers be able to accumulate information across situations that required the targeted skills or knowledge in terms of a student-model variable, which could then be used to trigger feedback, scaffolding, or instruction.

MashpeeQuest provides a second example. A central issue in any technology-based assessment is that of contextualization of skills to the technology being used. Often, exploiting the potential of technology—or of any material or social system for that matter—means learning about and taking advantage of its unique terminologies, conventions, and affordances. Indeed, from the point of view of situative psychology, contextualization is of the essence in learning:

> Knowing, in [the situative] perspective, is both an attribute of groups that carry out cooperative activities and an attribute of individuals who participate in the communities of which they are members. … Learning by a group or individual involves becoming attuned to constraints and affordances of material and social systems with which they interact. (Greeno, Collins, & Resnick, 1996, p. 17)

These insights challenge the familiar strategy of assessing through standardization—”measuring the same thing” for all students by gathering the same data under the same conditions. For example, AmericaQuest is intended to develop student skill in analyzing information and problem solving specifically in the context of an Internet-based adventure learning experience. The adventure involves using inquiry tools and evidentiary-reasoning skills typically used by historians and archaeologists, but in an important sense, the analysis and problem-solving skills students are learning are confounded with learning how to use the Internet to conduct inquiry. Observation data suggest, however, that teachers’ off-line instruction mediates students’ learning of these skills in significant ways (Penuel & Shear, 2000). If teachers’ assignments to students are unrelated to the central historical dilemma posed by the Quest, and students are not directed to weigh evidence about particular hypotheses, students will fail to learn (at least through AmericaQuest) the information analysis and problem-solving skills identified in the student model.

To what extent are the skills confounded with the technological environment? This question returns us to the issue of what we want to build into the student model—what we need to “tell stories about.” In the Classroom Connect evaluation plan, it was determined that some of the skills of interest could be found, to some degree, outside of the AmericaQuest technological environment. Other components of the evaluation plan are designed to provide evidence about them in ways that could be used as pretests, or as comparisons with students who are not familiar with the AmericaQuest technological environment. But this would be an incomplete evaluation, for evidencing some of the skills of interest depends on providing the environmental support and having had the students learn to exploit the affordances of this performance task. MashpeeQuest provides an opportunity to get direct evidence about these contextualized skills—but with different domain knowledge from the knowledge they encountered in AmericaQuest. We are thus attempting to define the skills students will obtain in a way that conditions on the technological environment but generalizes across the specifics of subject matter. This is evidence that cannot, by its very nature, be obtained from students who have not been acculturated in the AmericaQuest environment. Rather than obtaining a measure of skills that can be quantitatively compared with those of students from outside the program, MashpeeQuest provides evidence about the degree to which the AmericaQuest students exhibit skills they were meant to develop, in an environment in which their skills have been attuned. It provides evidence for a kind of "existence proof" story among the program students rather than a "horse race" story between these students and those from another program, or even between themselves before and after they experienced the program.

What we can observe to give us evidence. Given the terms in which we want to characterize students' capabilities, what can we observe that will constitute evidence of those capabilities? That is, what do we need to see in what a student actually says or does—the work product—and how do we characterize what we see when we see it—the evidence rules? Identifying what we need to see is especially important in complex performances. Even when we rely on largely empirical tools, such as neural networks, to evaluate key characteristics of a performance, success will depend on identifying the right kinds of features. For example, Stevens, Lopo and Wang (1996) produced neural nets that were better able to distinguish experts' diagnostic solutions from novices' solutions when they used sequenced pairs of the tests they ordered, rather than just which ones, as input features. There was less evidence about problem solving in the choice of tests that examinees performed than in which tests they performed after other tests; the experts were better able than novices to understand the implications of the results of one test to optimally select the next one.

Accumulating research in cognitive psychology again provides guideposts (e.g., Ericsson & Smith, 1991). What kinds of behaviors signal expert thinking? Similar patterns in the way experts think have been observed across many domains, as different as radiology is from volleyball or troubleshooting hydraulics systems is from solving middle-school electrical circuit problems. In general terms, experts…

> (a) provide coherent explanations based on underlying principles rather than descriptions of superficial features or single statements of fact, (b) generate a plan for solution that is guided by an adequate representation of the problem situation and possible procedures and outcomes, (c) implement solution strategies that reflect relevant goals and subgoals, and (d) monitor their actions and flexibly adjust their approach based on performance feedback. (Baxter, Elder, & Glaser, 1996, p. 133)

The trick is to understand the particular forms that these general patterns take in different domains. In the DISC project, researchers encoded them as "performance features." They identified these features from similarities in behaviors and reasoning across many problems from many hygienists at different levels of expertise. They needed to specialize to the representational forms, the problem environments and tools, and the knowledge structures and procedural requirements of the domain in question, but remain with statements sufficiently general to apply to many specific situations in that domain.

The kinds of historical-reasoning behaviors elicited in the MashpeeQuest example are behaviors that are parallel to the activities of professional historians. Expert historians spend much of their time analyzing historical texts, images, and artifacts (Wineburg, 1991), just as students in the MashpeeQuest task spent most of their time reading and interpreting the text on the various links to cities in the task. The MashpeeQuest scoring rubric would assign higher scores to student behaviors that suggested that students were not just spending time analyzing documents but also analyzing them in ways that are similar to the ways expert historians analyze documents (see Wineburg, 1998). Expert historians, for example, may consider how evidence in one document supports or contradicts evidence in another document, something that students are explicitly invited to consider in the MashpeeQuest task. Student skill in analyzing documents is made visible through the formulation of an argument backed by specific evidence from the documents, as well as a consideration of possible counter evidence from other links on the MashpeeQuest site.

Modeling which aspects of performance depend on which aspects of knowledge. The objective in the statistical model is expressing the ways in which certain aspects of performance depend on particular aspects of knowledge. As discussed above, the *purpose* of an assessment drives the number and granularity of student-model variables. But a CTA can additionally show how the skills and knowledge that tasks require are called on. An example from the HYDRIVE project illustrates the idea. HYDRIVE is a coached practice system for troubleshooting the hydraulics systems of the F-15 aircraft. The CTA (Steinberg & Gitomer, 1996) showed that not only are the elements of declarative, strategic, and procedural knowledge individually required for high probabilities of expert troubleshooting actions, but they are *all* required; lack of any of the three components impairs performance. The building block in the statistical model that expresses the relationship between this knowledge and successful troubleshooting steps is therefore conjunctive.

Effective ways to elicit the kinds of behavior we need to see. What characteristics of problems stimulate students to employ various aspects of their knowledge? We are beginning to hear phrases such as “principled task design” in assessment more often (e.g., Embretson, 1998). The idea is that by systematically manipulating the features of task settings—that is, controlling the constraints and the affordances—we create situations that encourage students to exercise targeted aspects of skill and knowledge. We describe these features in terms of task-model variables.

Work on systematic and theory-based task design dates back at least half a century. We may point to Louis Guttman’s (1959) facet design for tests, followed by Osburn’s (1968) and Hively, Patterson, and Page’s (1968) work in the 1960s with item forms and John Bormuth’s (1970) linguistic transformations of texts to produce comprehension items. But now we can take advantage of concepts and methods from psychology to build tasks more efficiently and around cognitively relevant—and therefore construct-relevant—features. We have discussed ways we can manipulate the medical conditions of patients and the availability of information in the DISC simulator environment, either to elicit evidence about hygienists' medical or information-gathering knowledge or to minimize the stress on this knowledge in order to highlight other aspects of their competence. We have also explored how a Web-based environment can be used to reveal student information analysis and problem-solving skills across a range of tasks; in particular, we have considered how the content available at different Internet links could be varied to isolate particular information analysis and problem-solving skills. A Web-based environment is a particularly adaptable vehicle for presenting assessment tasks. The wealth of information available on the Web makes varying the substance of the assessment task relatively easy, within an assessment schema under which task format and underlying targeted skills remain constant.

Leverage Points for Technology

Now let's look at some leverage points for using technology. We shall see that they can often be exploited to realize the possibilities that cognitive psychology offers.

Dynamic assembly of the student model. First is the capability to use contextual or concurrent information to bring up or assemble a student model. In interactive contexts, we can think of shifting the focus of our inquiry or switching the grain size of the student model as we learn about some parts of the model and update our options for action.

A simple example of this approach could be applied in the domain of document literacy (Kirsch & Jungeblut, 1986). An overall scale, from less proficient to more proficient, is useful when a potential student is referred to an adult literacy training program. It provides a quick idea of general level of proficiency, perhaps on the 100-500 scale of the National Adult Literacy Survey (NALS), for the purposes of documentation and program accountability. As an example, Meredith comes out at 200 on the scale. But further diagnostic assessment, focused for students in this same neighborhood of overall proficiency, is more useful for determining what to work on, because Meredith, Jessica, Bob, and seven other people at level 200 need different kinds of help to get to 250. Is Meredith familiar with the prototypical structures that documents are based on, such as lists, nested lists, and tables? What strategies does she have to work with? Does she recognize the kinds of situations that call for their use? Is vocabulary the stumbling block? Would help with reading be her best bet? What is key here is that the follow-up questions for students at level 200 are different from the follow-up questions for students at level 300, who want to get to 350. Tasks from the same pool as the initial assessment might be used for follow-up, but they would be hooked up with evidence models to inform finer-grained student models. The SM variables in these models would be tailored to feedback of different kinds for students at different levels of proficiency; they would be variables that answer a question like, "What is the *nature* of Meredith's proficiency, now that we know the *level* of her proficiency?"

Realistic tasks to produce direct evidence. Technology helps *designers* create complex and realistic tasks that can produce direct evidence about knowledge used for production and interaction. In part, this concerns capturing the richness and complexity of the environment we can create for the student, and in part it concerns the richness and complexity of the responses we can capture. Video capture of a dance, for example, requires no new technology for presentation, but it makes it possible for the ephemeral performance to be viewed and evaluated at many times and in many places—a wonderful mechanism for communicating evaluation standards (Wolf, Bixby, Glenn, & Gardner, 1991). The capacity of technology to capture an event does not just help improve the consistency of evaluation; it helps students learn about the standards of good work in the domain. This example illustrates an application of ideas from situative psychology: part of the social milieu of a student is participating in the assessment; the standards of evaluation are among the constraints of her environment; she must develop knowledge and skills to use the affordances of the settings to succeed in these socially required trials.

The MashpeeQuest performance assessment presents students with a realistic setting that they are likely to use on a regular basis in the 21st century to gather and evaluate information. MashpeeQuest requires students to be able to use the affordances of the Web-based environment to analyze text from multiple sources using a browser, and to use the Internet to communicate their ideas. It is not just the analysis skills that students learn, but the etiquette and protocol of communicating in the socially situated Internet community. Students' use of the Web-based learning environment is mediated, of course, by their classroom teacher's support, peer interactions and discussion, and their own skill in navigating the site. The MashpeeQuest assessment illustrates several of the ways in which technology can enhance the quality of assessment: it provides more possibilities in the content and formats that can be used to present materials and document students' competences, while at the same time providing task constraints to ensure that the assessment measures the construct it is intended to measure.

Automated extraction and evaluation of key features of complex work. Some automated extractions and evaluations of key features of complex work make it possible to increase the efficiency of applying existing evidence rules. Others make it possible to evaluate work products that could not be routinely include in assessment. In an example mentioned above, Stevens et al. (1996) used neural networks to summarize the import of students' sequences of diagnostic tests. Examples from current projects at Educational Testing Service include:

- Natural language processing methods for scoring essays, employing psycholinguistic and semantic theory used to define features to extract, and tree-based regression is used to summarize them into scores.

- Evaluation of constructed show-your-steps responses to algebra problems, with GOMS (Goals, Operators, Methods, and Selection Rules) methodology to infer students' likely strategies (Card, Moran & Newell, 1983).

- Automatic scoring of features of architectural designs, such as whether a student's floor plan gives enough space for a person in a wheelchair to get from the door to behind the desk, with automated routines to evaluate clearances along the person's path.

Examples from MashpeeQuest include:

- Counts of the number of Internet links checked and calculation of the amount of time spent examining each link.
- Evaluation of student reasoning by identifying whether evidence from particular links is used to support particular hypotheses.
- Comparison of students' own ratings of the relevance of particular links with experts' ratings.

Automated/assisted task construction, presentation, and management. In a preceding section, we discussed how research in cognitive psychology reveals systematic relationships between the affordances and constraints of problem situations and the knowledge structures and procedures people can bring to bear on those problems. Understanding and systematically manipulating these features of tasks not only helps us produce tasks more efficiently, it also strengthens the validity argument for them. Further benefits accrue if we can use technology to produce tasks as well. This is as true for producing familiar kinds of tasks as it is for ones that could not exist at all outside of a technological setting (such as DISC's computer-based simulations). Likewise, the Videodiscovery technology-based investigations and the SMART assessments developed by the Cognition and Technology Group at Vanderbilt illustrate the use of technology to assess phenomena that are too large, too small, too dynamic, too complex, or too dangerous to be validly assessed with non-technology-based methods of assessment (Vye Schwartz, Bransford, Barron, Zech, 1998). The production side of assessment can exploit technology in several ways, including, for example, automated and semiautomated construction of items (e.g., Bennett, 1999) and tools to create tasks according to cognitively motivated schemas (e.g., Embretson, 1998).

A further comment on technology-based assessment. Technology is as seductive as it is powerful. It is easy to spend all one's time and money designing realistic scenarios and gathering complex data, and only then to ask "How do we score it?" When this happens, the chances are great that the technology is not being used to best effect. The affordances and constraints are not selected optimally to focus attention on the skills and knowledge we care about and to minimize the impact of incidental skills and knowledge. Since affordances are not selected optimally, we emphasize the evidentiary foundation that must be laid if we are to make sense of any complex assessment data. The central issues are construct definition, forms of evidence, and situations that can provide evidence, regardless of the means by which data are to be gathered and evaluated. Technology provides such possibilities as simulation-based scenarios, but evidentiary considerations should shape the thousands of implementation decisions that arise in designing a technology-based assessment. These are the issues that cause such an assessment to succeed or to fail in serving its intended purpose. Messick's (1994) discussion on designing performance assessments is mandatory reading for anyone who wants to design a complex assessment, including computer-based simulations, portfolio assessments, and performance tasks.

In the case of DISC, the simulator needs to be able to create the task situations described in the task model and to capture that behavior in a form that DISC developers and researchers have determined they need to obtain evidence about targeted knowledge—that is, to produce the required work products. What possibilities, constraints, and affordances must be built into the simulator to provide the data needed? As to the kinds of situations that will evoke the behavior they want to see, the simulator must be able to:

- Present the distinct phases in the patient interaction cycle (assessment, treatment planning, treatment implementation, and evaluation).
- Present the forms of information that are typically used, and control their availability and accessibility, so we can learn about examinees' information-gathering skills.
- Manage cross-time cases, not just single visits, so we can get evidence about examinees' capabilities to evaluate information over time.
- Vary the virtual patient's state dynamically, so we can learn about examinees' ability to evaluate the outcomes of treatments that they choose.

As to the nature of affordances that must be provided, DISC has learned from the CTA that examinees should have the capacity to:

- Seek and gather data.
- Indicate hypotheses.
- Justify hypotheses with respect to cues.
- Justify actions with respect to hypotheses.

An important point is that DISC does not take the early version of the simulator as given and fixed. Ultimately, the simulator must be designed so the highest priority is providing evidence about the targeted skills and knowledge—not authenticity, not "look and feel," not technology.

As for MashpeeQuest, the assessment task situations must parallel the kinds of situations faced by students as they analyze information and solve problems in the AmericaQuest program, so that the assessment tasks are more likely to be sensitive to the effects of the program itself. It should capture student performances on skills both that are specific to AmericaQuest and those that are valued by educators and policy-makers who would look to the findings from an evaluation of AmericaQuest as the basis for decision-making about purchasing or continuing to use the program.

As to the kinds of situations that will evoke the behavior we want to see, the assessment must be able to:

- Present students with a historical or archaeological dilemma with competing hypotheses to consider.
- Present students with distinct phases of problem solving using historical documentation.

- Vary the problem or dilemma, to provide evidence for generalizability of student skills across tasks.
- Include multiple sources of pictorial and text-based evidence that can be used to support or to disconfirm different hypotheses.
- Allow for students to enter a text-based argument regarding their own position about the dilemma.
- Vary the outcomes of the search dynamically, so we can learn about students' ability to evaluate the outcomes of searches that they conduct.

In turn, the students being tested in this environment should be able to:

- Seek and gather data on the Internet.
- Carry out analyses of the evidence found on as many links as possible in the task.
- Construct a coherent argument in support of one hypothesis using evidence from the links, with both confirming and disconfirming evidence that can be discovered and taken into account.
- Enter text in an interactive Web-based environment setting.

Final Observations

These developments will have the most impact when assessments are built for well-defined purposes and connected with a conception of competence in the targeted domain. They will have much less impact for “drop-in-from-the-sky,” large-scale assessments like the National Assessment of Educational Progress. They are important in two ways for gauging students’ progress and evaluating the effectiveness of educational programs.

First, these developments can be exploited to design assessments that better home in on the most crucial questions in the application. But doing so requires resources—the time, energy, money, and expertise to tailor an assessment to a purpose. We expect that technologies coming available will continue to make it easier and cheaper to create more ambitious assessments and to share and tailor assessment building blocks that have been provided by others. For now, however, resources remain a serious constraint.

Second, then, in recognition of the limitations a lack of resources inevitably imposes, the new perspectives offered by developments in cognative psychology, technology, and evidence-centered assessment may be used today to select assessments from these currently available—to do as well as possible at focusing on what matters. Knowing how we would proceed with unlimited resources to create assessments that suited our purposes perfectly, we are in a better position to evaluate the quality of existing assessments we may have to choose among. We can better say what they tell us and what they miss—and perhaps save enough money to gather some supplementary data on just those facets of competence that off-the-shelf instruments cannot address.

References

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

Baxter, G. P., Elder, A. D., & Glaser, R. (1996). Knowledge-based cognition and performance assessment in the science classroom. *Educational Psychologist, 31*(2), 133-140.

Bennett, R. E. (1999). Using new technology to improve assessment. *Educational Measurement: Issues and Practice, 18*, 5-12.

Bormuth, J. R. (1970). *On the theory of achievement test items*. Chicago: University of Chicago Press.

Card, S., Moran, T., & Newell, A. (1983) *The psychology of human-computer interaction. Mahwah, NJ: Lawrence Erlbaum Associates*

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods, 3,* 380-396.

Ericsson, K. A., & Smith, J. (1991). Prospects and limits of the empirical study of expertise: An introduction. In K. A. Ericsson & J. Smith (Eds.), *Toward a general theory of expertise: Prospects and limits.* Cambridge: Cambridge University Press.

Glaser, R., Lesgold, A., & Lajoie, S. (1987). Toward a cognitive theory for the measurement of achievement. In R. Ronning, J. Glover, J.C. Conoley, & J. Witt (Eds.), *The influence of cognitive psychology on testing and measurement: The Buros-Nebraska Symposium on Measurement and Testing* (Vol. 3) (pp. 41-85). Hillsdale, NJ: Erlbaum.

Greeno, J. G., Collins, A. M., & Resnick, L. B. (1996). Cognition and learning. In D. Berliner & R. Calfee (Eds.), *Handbook of educational psychology* (pp. 15-47). New York: Simon & Schuster Macmillan.

Guttman, L. (1959). A structural theory for inter-group beliefs and action. *American Sociological Review, 24*, 318-328.

Hively, W., Patterson, H. L., & Page, S. H. (1968). A "universe-defined" system of arithmetic achievement tests. *Journal of Educational Measurement, 5*, 275-290.

Jensen, F.V. (1996). *An introduction to Bayesian networks*. New York: Springer-Verlag.

Johnson, L. A., Wohlgemuth, B., Cameron, C. A., Caughman, F., Koertge, T., Barna, J., & Schulz, J. (1998). Dental Interactive Simulations Corporation (DISC): Simulations for education, continuing education, and assessment. *Journal of Dental Education, 62*, 919-928.

Kirsch, I. S., & Jungeblut, A. (1986). *Literacy: Profiles of America's young adults.* Princeton, NJ: National Assessment of Educational Progress/Educational Testing Service.

Linn, M. C., Bell, P. & Hsi, S. (1999). Lifelong science learning on the Internet: The Knowledge Integration Environment. *Interactive Learning Environments, 6*(1-2), 4-38.

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., Almond, R. G., Yan, D., & Steinberg, L. S. (1999). Bayes nets in educational assessment: Where do the numbers come from? In K. B. Laskey & H. Prade (Eds.), *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence.* San Francisco: Morgan Kaufmann.

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., Steinberg, L. S., & Almond, R. G. (2002). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1,* 3-67.

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

Mislevy, R. J., Steinberg, L. S., Breyer, F. J., Almond, R. G., & Johnson, L. (2002). Making sense of data from complex assessment. *Applied Measurement in Education, 15,* 363-389.

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

Osburn, H. G. (1968). Item sampling for achievement testing. *Educational and Psychological Measurement, 28,* 95-104.

Penuel, W., & Shear, L. (2000). *Classroom Connect: Evaluation design.* Menlo Park, CA: SRI International.

Quellmalz, E., & Haertel, G. D. (1999). *Breaking the mold: Technology-based science assessment in the 21st century.* Menlo Park, CA: SRI International Technical Report.

Rogoff, B. (1984). Introduction. In B. Rogoff & J. Lave (Eds.), *Everyday cognition: Its development in social context* (pp. 1-8). Cambridge, MA: Harvard University Press.

Schum, D. A. (1994). *The evidential foundations of probabilistic reasoning.* New York: Wiley.

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

Stevens, R. H., Lopo, A. C., & Wang, P. (1996). Artificial neural networks can distinguish novice and expert strategies during complex problem solving*. Journal of the American Medical Informatics Association, 3,* 131-138.

Vye, N. J., Schwartz, D. L., Bransford, J. D., Barron, B. J., Zech, L., & The Cognition and Technology Group at Vanderbilt. (1998). SMART environments that support monitoring, reflection, and revision. In D. J. Hacker, J. Dunlosky, & A. C. Grasesser (Eds.), *Metacognition in educational theory and practice* (pp. 305-346). Hilldale, NJ: Erlbaum.

Wineburg, S. (1991). On the reading of historical texts: Notes on the breach between school and academy. *American Educational Research Journal, 28,* 495-519.

Wineburg, S. (1998). Reading Abraham Lincoln: An expert/expert study in the interpretation of historical texts. *Cognitive Science, 22,* 319-346.

Wolf, D., Bixby, J., Glenn, J., & Gardner, H. (1991). To use their minds well: Investigating new forms of student assessment. In G. Grant (Ed.), *Review of educational research* (Vol. 17) (pp. 31-74). Washington, DC: American Educational Research Association.

Sponsor

The National Science Foundation, Grant REC-0129331

Prime Grantee

SRI International. *Center for Technology in Learning*

Subgrantees

University of Maryland

University of California, Berkeley. *Berkeley Evaluation & Assessment Research (BEAR) Center* and *The Full Option Science System (FOSS)*

University of Michigan. *BioKIDs*