Making Sense of Data From Complex Assessments

**CSE Technical Report 538**

Robert J. Mislevy, Linda S. Steinberg, and Russell G. Almond
Educational Testing Service, Princeton, NJ
F. Jay Breyer
The Chauncey Group International, Princeton, NJ
Lynn Johnson
Dental Interactive Simulations Corporation, Aurora, CO

March 2001

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
301 GSE&IS, Box 951522
Los Angeles, CA 90095-1522
(310) 206-1532

Project 3.2 Validity of Interpretations and Reporting of Results—Evidence and Inference in Assessment Robert J. Mislevy, Project Director, CRESST/Educational Testing Service



The first author's work was supported in part by the Educational Research and Development Centers Program, PR/Award Number R305B60002, as administered by the Office of Educational Research and Improvement, U.S. Department of Education.

The findings and opinions expressed in this report do not reflect the positions or policies of the National Institute on Student Achievement, Curriculum, and Assessment, the Office of Educational Research and Improvement, or the U.S. Department of Education.

# MAKING SENSE OF DATA FROM COMPLEX ASSESSMENTS[1]

**Robert J. Mislevy, Linda S. Steinberg, and Russell G. Almond**
**Educational Testing Service, Princeton, New Jersey**

**F. Jay Breyer**
**The Chauncey Group International, Princeton, New Jersey**

**Lynn Johnson**
**Dental Interactive Simulations Corporation, Aurora, Colorado**

## Abstract

Advances in cognitive psychology deepen our understanding of how students gain and use knowledge, and broaden the range of performances and situations we want to see to acquire evidence about their developing knowledge. At the same time, advances in technology make it possible to capture more complex performances in assessment settings by including, as examples, simulation, interactivity, and extended responses. The challenge is making sense of the complex data that result. This presentation concerns an evidence-centered approach to the design and analysis of complex assessments. It presents a design framework that incorporates integrated structures for modeling knowledge and skills, designing tasks, and extracting and synthesizing evidence. The ideas are illustrated in the context of a project with the Dental Interactive Simulation Corporation (DISC) in which problem solving in dental hygiene is assessed with computer-based simulations. After reviewing the substantive grounding of this effort, we describe the design rationale, statistical and scoring models, and operational structures for the DISC assessment prototype.

Interest in complex and innovative assessment is increasing nowadays for a number of reasons. For one, researchers have opportunities to capitalize on recent advances in cognitive and educational psychology related to how people learn, how they organize knowledge, and how they put it to use (Greeno, Collins, & Resnick,

[1] This paper is based on research conducted for the Dental Interactive Simulation Corporation (DISC) by the Chauncey Group International (CGI), Educational Testing Service (ETS), and the DISC Scoring Team: Barry Wohlgemuth, DDS, DISC President and Project Director; Lynn Johnson, Ph.D., Project Manager; Gene Kramer, Ph.D.; and five core dental hygienist members, Phyllis Beemsterboer, RDH, Ed.D., Cheryl Cameron, RDH, Ph.D., JD, Ann Eshenaur, RDH, Ph.D., Karen Fulton, RDH, BS, and Lynn Ray, RDH, BS. Robert J. Mislevy currently is at the Department of Measurement, Statistics, and Evaluation, University of Maryland at College Park.

1997). This broadens the range of what we want to know about students, and what we might see 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).

But how are we to make sense of data from complex assessments? Don Melnick, who for several years led the National Board of Medical Examiners (NBME) project on computer-based case management problems, observed, "The NBME has consistently found the challenges in the development of innovative testing methods to lie primarily in the scoring arena. Complex test stimuli result in complex responses which require complex models to capture and appropriately combine information from the test to create a valid score" (1996, p. 117). The statistical methods and rules-of-thumb that evolved to manage classroom quizzes and standardized tests often fall short of this goal.

This presentation is based on two premises. The first is that the tools of probability-based reasoning, which specialize to familiar test theory for modeling data from familiar forms of assessment testing, can be applied from first principles to model complex data from innovative forms of testing (Mislevy, 1994; Mislevy & Gitomer, 1996). Recent developments in statistics and expert systems make it possible to build models and obtain estimates for more complex situations than were hitherto possible. The second premise is that flexible models and powerful statistical methods alone aren't good enough. It is a poor strategy to hope to figure out "how to score it" only after an assessment has been constructed and performances have been captured. A better approach would be to design a complex assessment from the very start around the inferences one wants to make, the observations one needs to ground them, the situations that will evoke those observations, and the chain of reasoning that connects them.

To this end, we have been developing an "evidence-centered" framework for designing assessments, as part of a project called PORTAL (Mislevy, Steinberg, & Almond, in press). We are using this framework to tackle design and scoring issues for a simulation-based assessment of problem solving in dental hygiene, a project of the Dental Interactive Simulations Corporation, or DISC (Johnson et al., 1998). A previous paper (Mislevy, Almond, Yan, & Steinberg, 1999) described the cognitive task analysis that was carried out to provide substantive and psychological grounding for the proposed assessment. This presentation concerns the next stage of the project, namely, constructing the design objects around which an operational

assessment can be built. The first part of the presentation reviews the design framework and the cognitive task analysis. The next part describes how the design objects were fleshed out. Particular emphasis is placed on evidence models—reusable substantive and statistical structures that frame the evidentiary argument from observations of complex data to inferences about complex skills. Finally, the ways that the pieces are assembled for assessing examinees in an operational program are outlined.

## Reasoning from Complex Data

So how should we make sense of data from complex assessments? We may begin by asking more generally how people make sense of complex data. Just how do we reason from masses of data of different kinds, fraught with dependencies, hiding redundancies and contradictions, each addressing different strands of a tangled web of interrelationships? Put simply, humans interpret complex data in terms of some underlying "story." It might be a narrative, an organizing theory, a statistical model, or some combination of many of these. It might be a simplifying schema we can hold in mind all at once, such as the verse "Thirty days hath September..." that helps us remember how many days each month has, or a complicated structure, such as a compendium of blueprints for a skyscraper. This is how we reason in law, in medicine, in weather forecasting, in everyday life (Schum, 1994). We weave some sensible and defensible story around the specifics. Such a story addresses what we really care about at a higher level of generality and a more basic level of concern than any of the particulars. A story builds around what we believe to be the fundamental principles and patterns in the domain.

In law, for example, every case is unique, but the principles of reasoning and principles for building stories are common. Jurists use statutes, precedents, and recurring themes from the human experience as building blocks to understand each new case (Pennington & Hastie, 1991). And one might characterize science as a principled approach to creating and checking stories. Research in cognitive decision making suggests that people are "wired" with certain patterns that they use as building blocks for reasoning—heuristics, such as estimating prevalence from familiarity and causation from co-occurrence. While such heuristics generally are useful, sometimes they are dead wrong (Kahneman, Slovic, & Tversky, 1982). Gardner (1991) argues that in any discipline, building blocks derived from

principled understandings of "the way things really work" are hard won for just this reason.

Equally important are the building blocks of evidentiary reasoning we must use to connect what we know about a substantive domain with what we see in the real world. Insights into evidentiary reasoning in a general form—that is, patterns and principles that apply across many domains, each working with its own underlying substance and forms of evidence—have appeared over the years in fields as varied as philosophy, jurisprudence, statistics, and computer science. The approach we take here can be viewed as the application of the general approach espoused by Schum (1987, 1994): structuring arguments from evidence to inference in terms of generating principles in the domain, and using probability-based reasoning to manage uncertainty. (Kadane and Schum, 1996, illustrate this approach in a fascinating analysis of the evidence from the famous Sacco and Vanzetti trial.)

## Evidentiary Reasoning and Assessment Design

In educational assessment, the building blocks of the stories that connect what students know and can do with what students say and actually do come from the nature of reasoning in assessment and the nature of the learning in the domain in question. The previously mentioned PORTAL project provides a conceptual framework and supporting software tools for designing assessments in this light. The project has three distinguishable aspects: (a) an evidence-centered perspective on assessment design, (b) object definitions and data structures for assessment elements and their interrelationships, and (c) integrated software tools to support design and implementation. In this presentation, we draw upon the perspective and a high-level description of the central objects and interrelationships. This section sets the stage by laying out the basic structure of the PORTAL conceptual assessment framework, or CAF. Then we discuss in greater detail how the ideas play out in the DISC prototype assessment.

Figure 1 is a high-level schematic of the three basic models we suggest must be present and must be coordinated to achieve a coherent assessment. A quote from Messick (1994, p. 17) serves well to introduce them:

> A construct-centered approach [to assessment design] would begin by asking what complex of knowledge, skills, or other attribute 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.

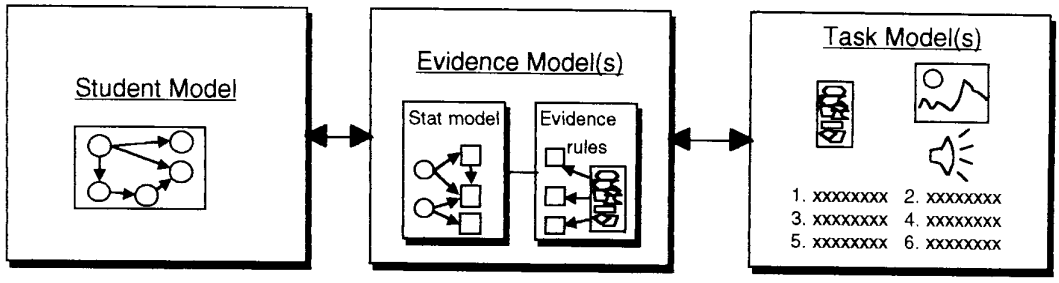



*Figure 1.* Three basic models of assessment design.

## The Student Model

"What complex of knowledge, skills, or other attributes should be assessed?" This is what the student model is about. Configurations of values of student-model variables are meant to approximate certain aspects of the infinite configurations of skill and knowledge real students have, as seen from some perspective about skill and knowledge in the domain. It could be the perspective of behaviorist, trait, cognitive, or situative psychology. This perspective determines the kinds of stories we want to weave for our purposes, but the evidentiary problem of constructing them from limited evidence is essentially the same. These are the terms in which we want to talk about students—the level at which we build our story 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 construe that as evidence about the student-model variables. In addition to one's conception of competence in the domain, the number and nature of the student-model variables in a particular application also depend on the purpose of the assessment. A single variable characterizing overall proficiency might suffice in an assessment meant to support only a summary pass/fail decision. A coached practice system that helps students develop the same proficiency would require a finer grained student model for monitoring how a student is doing on particular aspects of skill and knowledge for which we can provide feedback.

The student model in Figure 1 depicts student-model variables as circles. The arrows connecting them represent important empirical or theoretical associations

among them. We will use a statistical model to manage our knowledge about a given student's unobservable values for these variables at any given point in time, expressing it as a probability distribution that 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 (see Jensen, 1996, for an introduction to Bayes net from a statistical perspective; Edwards, 1998, for a modeling perspective; and Mislevy, 1994, and Almond & Mislevy, 1999, for an assessment perspective). Appendix A gives a simple example of how one can use the framework of Bayes nets to manage knowledge and uncertainty in assessment. We will look more closely at their use in the DISC example covered later.

**Evidence Models**

"What behaviors or performances should reveal those constructs," and what is the connection? This is what evidence models are about. An evidence model is the heart of evidentiary reasoning in assessment. Here's where we lay out our argument about why and how our observations in a given task situation constitute evidence about student-model variables.

Figure 1 shows that there are two parts to the evidence model. The evaluative submodel concerns extracting the salient features of whatever the student says, does, or creates in the task situation, that is, the work product. The work product is represented by a rectangle containing a jumble of complicated figures at the far right of the evidence model. It is a unique human production, perhaps as simple as a mark on a machine-readable answer sheet, perhaps as complex as repeated cycles of evaluation and treatment in a patient-management problem. Three squares are shown coming out of the work product. They represent observable variables, or evaluative summaries of whatever the assessment designer has determined are the key aspects of the performance to take away as nuggets of evidence. The evaluative rules describe how to carry out these mappings from unique human actions into a common interpretative framework. This is where one lays out the argument about what is important in a performance, in light of the purpose of the assessment. These mappings can be as simple as determining whether the mark on a multiple-choice answer sheet is the correct answer, or as complex as an expert's holistic evaluation of four key aspects of an unconstrained patient-management solution. They can be automatic or they can require human judgment, or some combination of both.

The statistical submodel of the evidence model expresses how the observable variables depend, in probability, on student-model variables. This is where one lays out the argument for synthesizing evidence across multiple tasks or from different performances. Figure 1 shows that the observables are modeled as depending on some subset of the student-model variables. Familiar models from test theory are examples of statistical models in which values of observed variables depend probabilistically on values of unobservable variables. Among these are classical test theory, item response theory, latent class models, and factor analysis. More generally, we can express these familiar models 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).

### Task Models

"What tasks or situations should elicit those behaviors?" This is what task models are about. A task model provides a framework for constructing and describing the situations in which examinees act. Task-model variables play many roles in assessment, including structuring task construction, focusing the evidentiary value of tasks, guiding assessment assembly, implicitly defining student-model variables, and conditioning the statistical argument between observations and student-model variables (Mislevy, Steinberg, & Almond, in press). A task model includes specifications for the environment in which the student will say, do, or produce something. Examples include characteristics of stimulus material, instructions, help, tools, and affordances. A task model also includes specifications for the work product—the form in which what the student says, does, or produces will be captured. The data from a given task cannot be analyzed with a given evidence model unless the specifications of the work product for tasks written under the corresponding task model agree with the specifications of the work product that the evidence model expects.

## The DISC Project

In 1990, a consortium of dental education, licensure, and professional organizations created the Dental Interactive Simulation Corporation to develop computerized assessments and continuing-education products that simulate the work dentists and dental hygienists perform in practice (Johnson et al., 1998). The consortium directed DISC to develop as an initial application a computer-based

performance assessment of problem solving and decision making in dental hygiene. This assessment would fill a gap in the current licensure sequence. Hygienists provide preventive and therapeutic dental hygiene services, including educating patients about oral hygiene; examining the head, neck, and mouth; and performing prophylaxes, scaling, and root planing. Currently, multiple-choice examinations probe hygienists' content knowledge as it is required in these roles, and clinical examinations assess their skill in carrying out the procedures. But neither form of assessment provides direct evidence about the processes that unfold as hygienists interact with patients: seeking and integrating information from multiple sources, planning dental hygiene treatments accordingly, evaluating the results over time, and modifying treatment plans in light of outcomes or new information.

As this paper is written, DISC has developed a prototype of a dental simulation system in the context of continuing education (Johnson et al., op cit.). The simulator uses information from a virtual-patient database as a candidate works through a case. Some of the information is presented directly to the candidate (e.g., a medical history questionnaire). Other information may be presented on request (e.g., radiographs at a given point in the case). Still other information is used to compute patient status dynamically as a function of the candidate's actions and the patient's etiology. A student can thus work through interactions with virtual patients—gathering information, planning and carrying out treatments, and evaluating their effectiveness. These capabilities provide a starting point for the proposed simulation-based dental hygiene licensure assessment.

Educational Testing Service, under a subcontract with the Chauncey Group International, is working with DISC to develop a scoring engine for the proposed prototype of a simulation-based assessment of problem solving in dental hygiene. As previously stated, we believe this requires a broader perspective than just looking for a statistical model to make sense of whatever data happens to appear from whatever tasks and interfaces happen to have been produced. We are thus working through student, evidence, and task models with DISC, and in turn, examining the implications for the simulation system. These CAF models provide the foundation upon which we assemble the building blocks of our evidentiary arguments. The substance of the arguments concerns the nature of knowledge in dental hygiene, how it can be evidenced, and what is needed to serve the purpose of the DISC assessment.

The following section sketches the assessment framework we have developed for the proposed assessment, showing how its elements depend on, and are derived from, the models of the CAF. The section after that reviews substantive grounding of the project, including the cognitive task. Further discussion of the structure and contents of the design objects then follows.

## Design Rationale

### The Cognitive Task Analysis

A group of dental-hygiene experts assembled by DISC to serve as the DISC Scoring Team began by mapping out the roles and contexts of the work that dental hygienists perform, drawing on curricular materials, research literature, existing licensure tests, and personal experience. These materials also constituted a compendium of declarative knowledge that could be drawn upon both in the design of the cognitive task analysis (CTA), as described later, and in subsequent developmental work, such as defining task-model variables and their values.

A traditional job analysis focuses on valued tasks in a domain in terms of how often people must perform them and how important they are. A cognitive task analysis, in contrast, focuses on the knowledge people draw upon to carry out those tasks. A CTA in a given domain seeks to shed light on (a) essential features of the situations, (b) internal representations of situations, (c) the relationship between problem-solving behavior and internal representation, (d) how the problems are solved, and (e) what makes problems difficult (Newell & Simon, 1972). With creating assessment structures as the ultimate objective, we adapted CTA methods from the expertise literature (Ericsson & Smith, 1991) to capture and to analyze the performance of hygienists at different levels of expertise under standard conditions across a range of valued tasks. Details of the CTA appear in Mislevy, Steinberg, Breyer, Almond, and Johnson (1999). The work can be summarized as follows.

Working from the compendium of resources, the DISC Scoring Team created nine representative cases that require decision making and problem solving and would be likely to elicit different behavior from hygienists at different levels of proficiency. To produce stimulus materials for the cases, the team began with blank dental forms and charts commonly found in oral health care settings, and a corpus of oral photographs, enlarged radiographs, and dental charts of anonymous patients.

We gathered talk-aloud solutions from 31 subjects. There were approximately 10 at each of three levels of expertise: novices, or students midway through training; competent hygienists who recently received their licenses; and acknowledged experts, each of whom has had several years of experience in practice and most of whom also teach in dental education programs. A subject, a scoring team member, and one or two psychologist researchers participated in each interview. The scoring team member provided the brief verbal description of the patient in each case in turn. The researcher asked the subject to describe her thoughts out loud and describe what she would do next. As the subject progressed through the case, she would call for printed information, ask questions, and make assessment, treatment, patient education, and evaluation decisions. With each action, the scoring team member provided responses in the form of medical or dental history charts, radiographic, photographic, or graphic representations when available, or verbal descriptions of what the patient would say or what the result of a procedure would be. The researcher would ask the subject to interpret the information; for example, the hygienist would be asked to verbalize her thoughts in reaction to the stimulus, what it might mean, what hypotheses it might have sparked, or which further procedures it might indicate. The interviewers did not give feedback as to the underlying etiology of a case, the appropriateness of the subject's actions, or the accuracy of her responses. The case continued until the presenting problem was resolved.

### Performance Features

The DISC Scoring Team's mission was to abstract general characterizations of behavior patterns from the unique and specific actions of 31 subjects in nine particular cases. The team sought 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. In line with the goal of assessment, the team sought patterns that would be useful in distinguishing hygienists at different levels of competence. We refer to the resulting characterizations as "performance features" (see Figure 2). Examples include:

- **Using disparate sources of information.** Novice hygienists usually were 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.

- **Scripting behavior**. Novice hygienists often followed standard scripts to work through initial assessments and patient education, while more experienced hygienists increasingly tailored their actions to the conditions and characteristics of specific patients.

- **Investigation of hypotheses**. Expert performance generally was characterized by pursuing information where it leads. As in many other domains reported in the expertise literature, "... some of the protocols show [novices] less able to efficiently modify a schema in response to new data, in contrast to the experts, who were flexibly opportunistic, neither too fixated nor uncontrollably labile" (Lesgold et al., 1988, p. 319). Competent dental hygienists often only partially investige hypotheses. Novice dental hygienists frequently do not investigate hypotheses. If they do recognize that a problem exists, they may ask another professional to investigate it (not a bad thing in itself!) at an earlier stage than more experienced hygienists, who would gather further information in order to refine, confirm, or disconfirm early hypotheses.

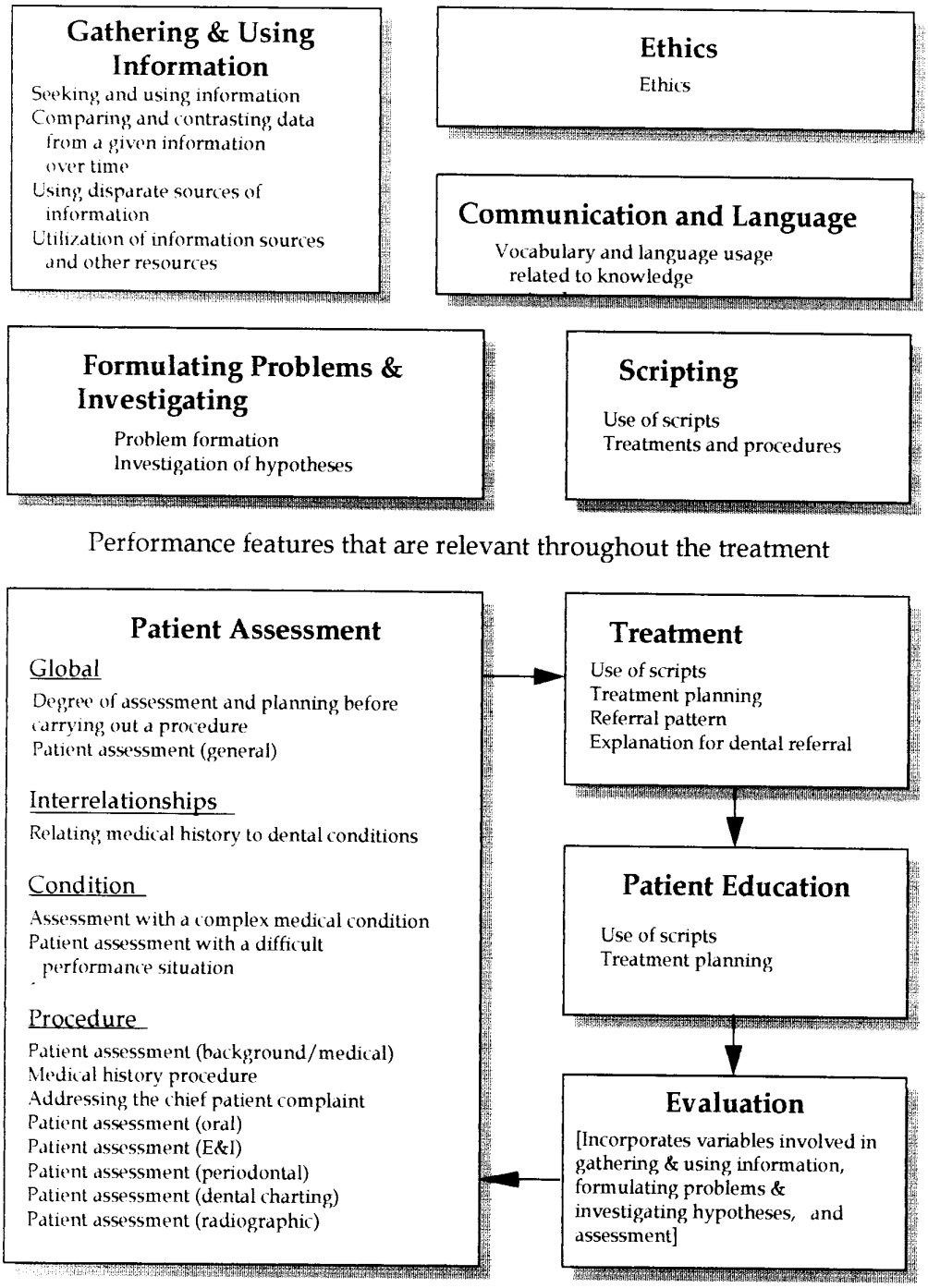



*Figure 2.* Performance features in decision making and problem solving in dental hygiene.

## Design Objects

The patterns summarized as performance features are cognitively grounded indicators of developing expertise in the domain of dental hygiene (Glaser, Lesgold, & Lajoie, 1987). We want to build the assessment around this collection of indicators of expertise. What are their implications for constructing the design objects, namely, student, evidence, and task models?

### The Student Model

The key consideration for determining a student model is the student-model variables' consistency with both the results of the CTA and the purpose of the assessment. In the case of the DISC prototype assessment, we ask more specifically: What aspects of skill and knowledge might be used to accumulate evidence across tasks, to summarize for pass/fail reporting, and to offer finer-grained feedback? We drew up a number of possible models. Figure 3 is a simplified version of the model we are using at present.

Figure 3 depicts a Bayes net that contains just the student-model variables. We shall refer to it as the student-model Bayes net fragment, because it will be combined with other fragments that include observable variables when observations become available and we want to update our beliefs about the student-model variables. The student-model variables in Figure 3 are represented as ovals. As examples, two ovals toward the upper right are *Assessment*, which concerns proficiency in assessing the status of a new patient, and *Information gathering/Usage*, which concerns proficiency in gathering and using information about patients. The full model, not shown, further elaborates the second of these. Finer grained student-model variables that are part of *Information gathering/Usage* include knowing how and where to obtain information, being able to generate hypotheses that would guide searches and interpretations, and knowing how to gather information which would help confirm or refute hypotheses.

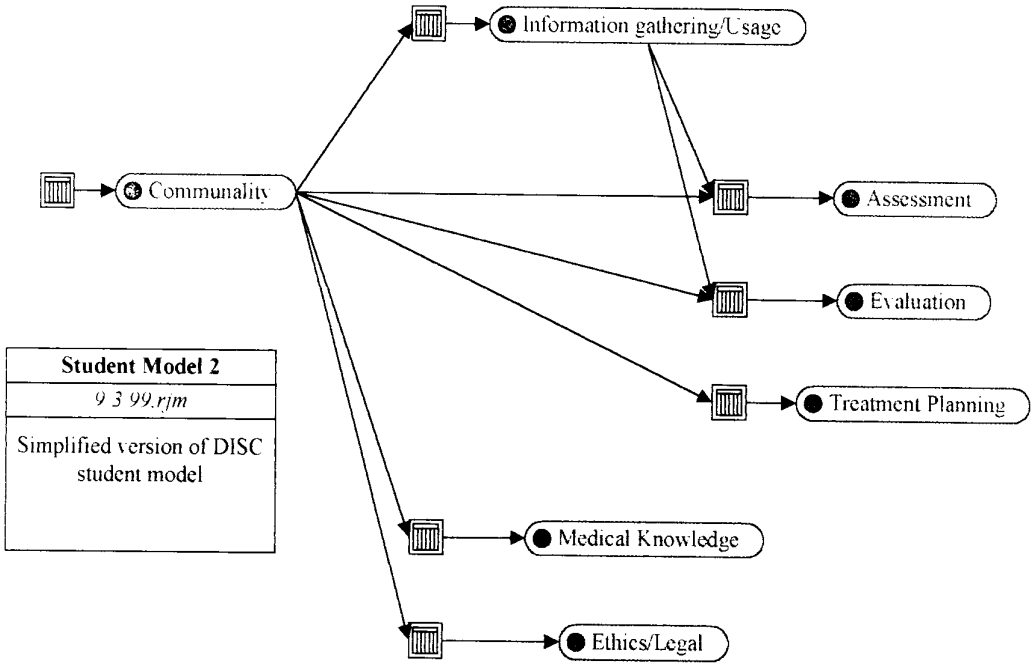



*Figure 3.* Simplified DISC Student Model.

Each variable has an associated square that represents the probability distribution that expresses current belief about a student's unobservable value on that variable. In this example, each variable is defined as having three levels, low, medium, and high, which correspond to novice, competent, and expert status. The arrows indicate associations among the variables; specifically, they represent conditional dependence relationships. A hygienist who has high proficiency in gathering and using information, for example, is more likely to have high or at least medium proficiency in assessing the initial status of new patients. The *Communality* student-model variable is a mechanism for incorporating our expectation that all of the more finely grained aspects of dental-hygiene proficiency may be correlated in a population of interest. Even if our interest lies at the more detailed level, allowing for associations serves two useful purposes. First, we will be able to exploit direct evidence about one aspect as indirect evidence about another. Second, if DISC wishes to project the finer grained student-model variables to a summary scale for a pass/fail decision, modeling their associations permits us to calculate measures of accuracy of such functions of student-model variables that correctly take the dependencies among the variables into account.

At the beginning of an examinee's assessment, the probability distributions representing a new student's status will be relatively uninformative (perhaps an

empirical estimate of the distribution in a population to which the examinee belongs, or perhaps a very diffuse prior) so that the final probability distribution will reflect evidence from her actions almost exclusively. The following sections will discuss how we successively update the joint distribution of the student-model variables to reflect our evolving belief as we make observations. Evidence models provide the technical machinery for making these changes in accordance with the evidentiary argument that justifies them.

**Evidence Models**

The student-model variables represent the proficiencies in which our interest lies, but they are inherently unobservable. Appendix A shows with a simple example how one uses Bayes nets to update beliefs about unobservable proficiency variables using evidence in the form of values of observable variables. What might be useful observable variables in this dental hygiene application? And what might be prototypical structures for getting evidence in these forms to tell us about students' proficiencies—structures around which many individual cases can be built?

The CTA produced performance features that characterize patterns of behavior and differentiate levels of expertise. They are grist for defining generally defined, reusable observed variables in evidence models. The evidence models themselves are structured assemblies of student-model variables and observable variables, including methods for determining the values of the observable variables and updating student-model variables accordingly. Appendix B provides a list of the 33 reusable evidence models we defined for use with potential DISC cases. As is described, a particular case will utilize the structures of one or more evidence models, fleshed out in accordance with specifics of that case.

Figure 4 depicts the Bayes net fragment that comprises the statistical submodel of one particular evidence model we'll use to discuss the building-block aspect of evidentiary reasoning. It concerns gathering patient information when assessing a new patient's status in the absence of inherent ethical or medical complications. We'll use it to show how evidence models are built from reusable structural elements, then tailored to the specifics of individual cases.

At the far left are student-model variables we posit to drive performance in these situations: *Assessment* of new patients and *Information gathering/Usage*, the two that were highlighted in the previous student model discussion. The *Context*

variable at the lower left accounts for dependencies among different aspects of performance in the same setting in order to avoid double-counting evidence that arises as different aspects of the same performance. The nodes on the right are generally defined observable variables. One, for example, refers to how well the examinee succeeds at *adapting to situational constraints.* Another refers to the *adequacy of examination procedures* in terms of how well their rationale is grounded. All the observable variables are defined as having between two and five possible values, generally ordered from poor to high quality. *Adequacy of examination procedures,* for example, has three values: *All* of the necessary points of an appropriate rationale are present in the examinee's solution, *some* are present, or *none* or few are present. These are generic categories that will be particularized for actual specific cases, the part of the evidentiary argument that is addressed in the evaluation submodel.

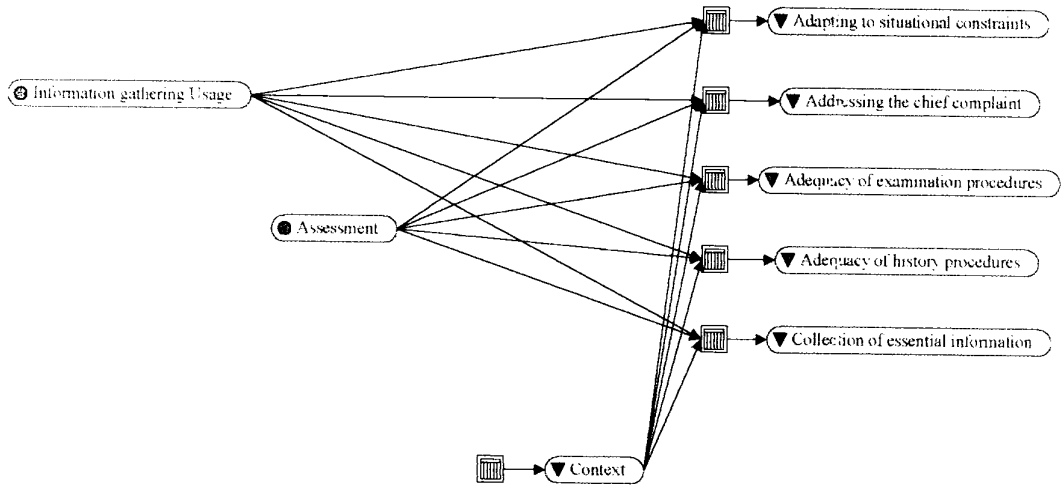



*Figure 4.* The Bayes net fragment in an evidence model.

## The Evaluation Submodel

As mentioned earlier, the evaluation submodel of an evidence model concerns the mappings from unique human actions or productions into a common framework of evaluation; that is, from work products to values of observable variables. Many tasks can be built around the same evidence model because the structure of the evidentiary argument is essentially the same.

What is constant in the evaluation submodels for tasks that are built to conform to the same evidence model are (a) the identification and formal definition of

observable variables, and (b) generally stated "proto-rules" for evaluating their values. *Adequacy of examination procedures* 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 in terms of, say, an ordered variable with the following three values: high, or adequate grounding for examination procedures; medium, or partial grounding of procedures; and low, or inadequate grounding.

What is customized to particular cases are rules for evaluating values of observables—tailored instantiations of the proto-rules that address the specifics of a 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 assessment and why. Experts must then specify the features of the content and rationale of examinees' assessment procedures that will determine the mapping to high, medium, and low values for this observable variable.

The "Mr. Clay" case, for example, requires gathering information to assess the status of a new patient. Given the specifics of the setup and the information about Mr. Clay, experts determined that an examinee should base examination procedures on two grounds: his chief complaint and the exception items on his health history review. A rationale having both grounds gets mapped to the high value; a rationale with just one gets mapped to medium; and a rationale with neither gets mapped to the low value.

### The Statistical Submodel

The statistical submodel of an evidence model concerns the synthesis of evidence from multiple or different tasks (arriving in the form of values of observable variables), in terms of our evolving beliefs about student model variables. It consists of the structure and the conditional probabilities in a Bayes net fragment, as seen in Figure 4.

What is constant in the statistical submodels of tasks that are built around the same evidence model are (a) the identities of the student-model parents, (b) the identities of the observable variables, and (c) the structure of the conditional probability relationships between the student-model parents and their observable children. For example, both proficiency in *Information gathering/Usage* and proficiency in *Assessment* of new patients are required in order to have high

probabilities of adequately grounding assessment procedures in any case that involves assessing a new patient.

What is customized to particular cases are (a) the specific meanings of the observables through the case-specific evaluation rules previously discussed, and (b) the values of the conditional probabilities that specify how potential outcomes depend on the values of student-model variables. Are the constraints imposed for this Virtual Patient *A* quite straightforward, for example, so even novices are likely to adapt to them? Are the constraints for Virtual Patient *B* subtle but demanding, so that even experts are not likely to make all of the ideal accommodations? The values of the conditional probabilities can be approximated initially from expert opinion and knowledge about the specific features of the task. Empirical data can be used to refine the estimates, much as we estimate the parameters in item response theory models (Mislevy, Almond, Yan, & Steinberg, 1999).

## Accumulating Evidence

Student-model variables and observable variables play asymmetric roles when we assess an examinee. Our interest in the (unobservable) values of the examinee's student-model variables is persistent. We want to make decisions and provide feedback based on their values as we learn about them from the examinee's performance across a series of tasks. The values of the observables her task performances produce are of interest only insofar as they allow us to update our beliefs about her student-model variables.[2] This is what happens during the course of observation: We start with a student model with probability distributions that indicate we know very little about this new examinee. We administer a case, and as the examinee works through the phases of the encounter, we successively "dock" a sequence of appropriate evidence models to incorporate information from her performances (Almond, Herskovits, Mislevy, & Steinberg, 1999).

The evidence model we've been considering has two student-model variables we posited to drive probabilities of actions in a certain class of situations. These variables are the link between the evidence-model Bayes net fragment and the student-model Bayes net fragment. Encountering a task situation for which this

[2] However, in a coached practice system or a self-diagnosis program, the values of observable variables can trigger "local" feedback—that is, insights or comments based on an examinee's particular action that are useful regardless of current belief about her status on the student-model variables.

evidence model is appropriate, we construct a combined Bayes net from the student-model Bayes net fragment and the evidence-model Bayes net fragment (Figure 5). We parse the work product and evaluate the observed variables. We enter these values in this combined Bayes net and update our beliefs about the student-model variables.

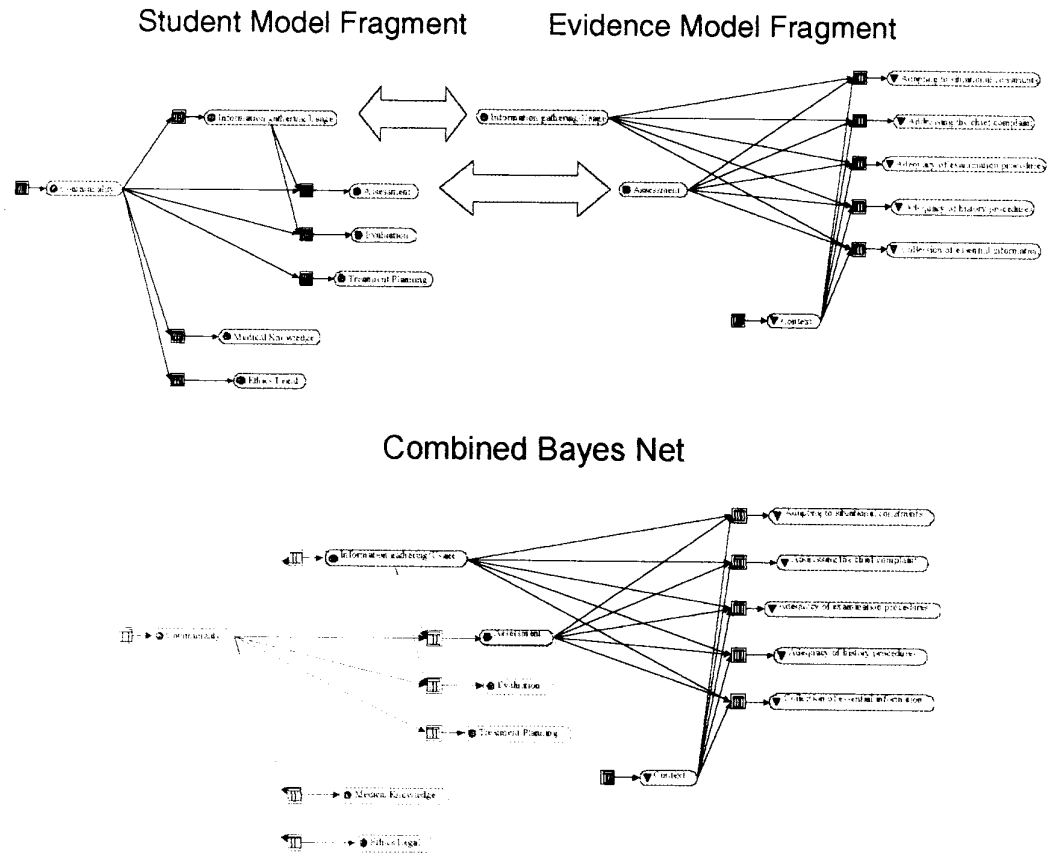



*Figure 5.* Docking student-model and evidence-model Bayes net fragments.

Figures 6 through 8 provide numerical examples, using a representation that shows probability distributions for each variable. Figure 6 is the initial setup: The student model, at initial conditions, with the Bayes net from the evidence model docked onto it, but with no observations made yet. Figure 7 is how our belief would change if we observed high-quality values on four observable variables that can be evaluated from the work product. The values for one of them are shown in the box at the right; they show a probability of one for the value that was actually observed and zero for the other two. The probabilities in the student-model variable for

*Assessment* proficiency have, appropriately, shifted upward. Figure 8 shows how the network would look if instead we obtained one high-quality and three low-quality values for the observable variables. This time the probabilities in the student model variable for *Assessment* proficiency have shifted downward. Having made these observations and registered their impact on our beliefs, we can jettison the Bayes net fragment for this evidence model—keeping the student-model fragment, with its updated probability distributions—and move to the next phase of the case, or to a different case, or stop testing and report results.

The DISC scoring engine allows DISC to specify arbitrary functions of student-model variables in order to obtain summary scores and accompanying measures of precision that are based on the final or any intermediate state of the student-model distribution. The "Next Steps" section toward the end of this presentation has a bit more to say about these functions.

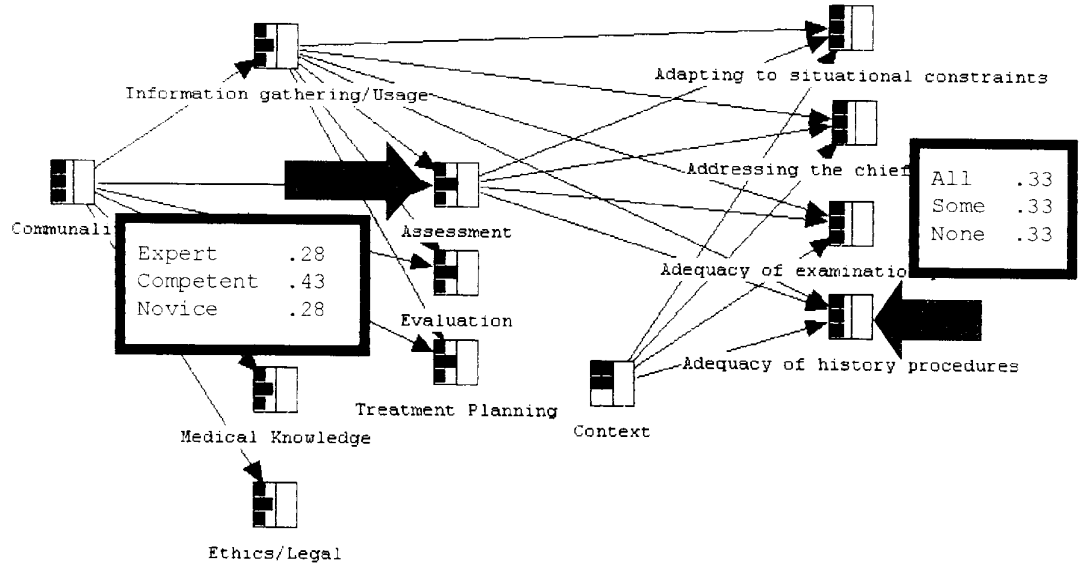



*Figure 6.* Numerical representations of initial status.

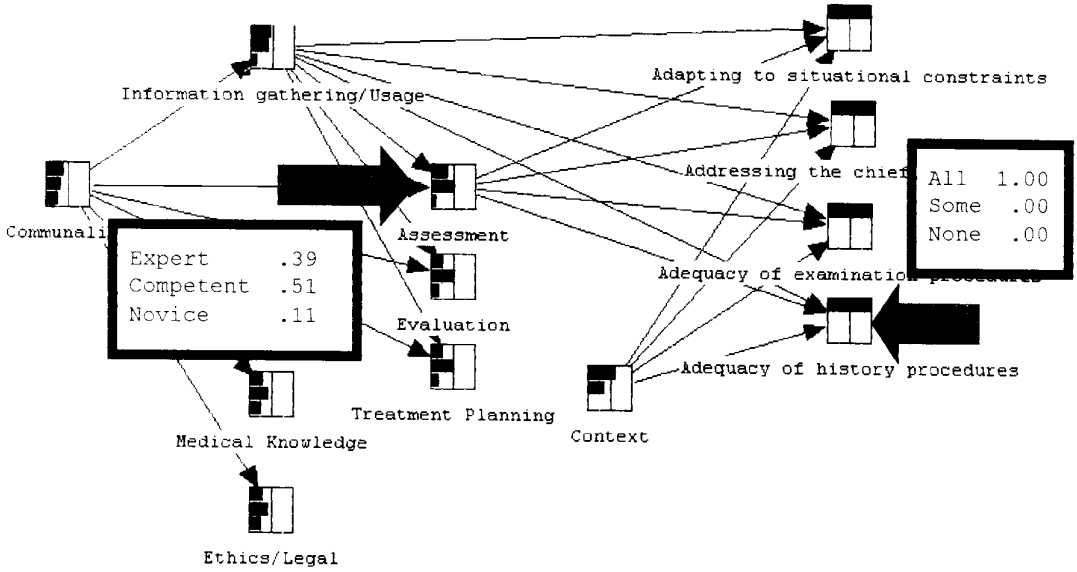



*Figure 7.* Numerical representations of status after four "good" findings.

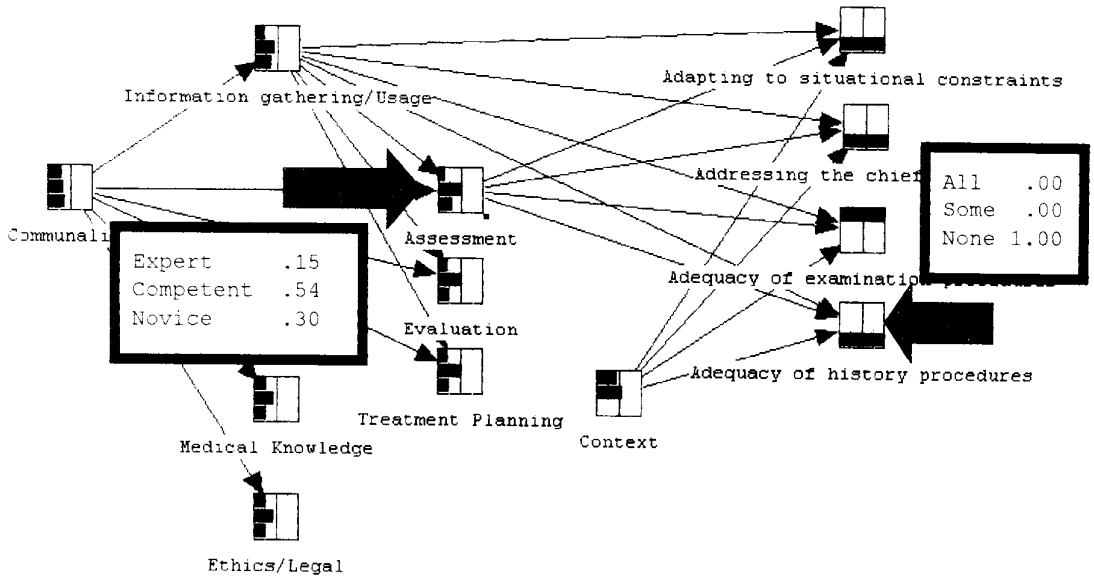



*Figure 8.* Numerical representations of status after one "good" and three "bad" findings.

## Task Models

Task models are the building blocks for the situations in which we make observations. Task models, it will be recalled, are schemas for describing and creating the situations that evoke the actions we need to see, so we'll be able to determine values for those observable variables in the evidence models. For the DISC prototype, we need to define task-model variables that

- The simulator needs for the virtual-patient database.
- Characterize features we need to evoke specified aspects of skill and knowledge.
- Characterize features of tasks that affect their difficulty.
- Characterize features we'll need to assemble tasks into tests.

This section gives examples of task-model variables with an eye toward these functions.

To anticipate one of the key ideas from the "Putting the Pieces Together" section coming up later, a test developer can create a case by first referring to the Proficiency/Evidence/Task (P/E/T) matrix. The P/E/T matrix is a cross-referenced list of the student-model variables, evidence models that can be used to get

information about each of the student-model variables, and task models around which tasks can be constructed that furnish values for the observables in those evidence models. Once a task model is selected, it is fleshed out with particulars to create a new virtual patient. The values of all the task-model variables used in that task model are determined accordingly.

There are distinguishable groups of task-model variables, many of which are hierarchically related. Here are two examples, using some of the task-model variables the scoring team proposed:

- *Oral Hygiene Status* is a task-model variable with possible values {excellent, good, poor, extremely poor}. Another task-model variable that is part of *Oral Hygiene Status* is *Bacterial Plaque*, which has possible values {Stage 1, Stage 2, Stage 3, heavy, moderate, light, none}; *Plaque Location* is part of *Bacterial Plaque* in turn, and indicates where the plaque, if it is present, is found.
- *Periodontal Status* is a very important task-model variable. Nested within it are the finer-grained task-model variables *Gingival Status-Attached* and *Gingival Status-Marginal*, and within each of them, still finer-grained task-model variables for *Color*, *Probing Depth*, *Contour*, and *Size*. Values for the most detailed level include both "within normal limits" (WNL) and various classes of irregularities that can be used as cues for underlying etiologies. The presence and severity of values other than WNL determine values of the higher level variables such as *Periodontal Status*.

Examples of task-model variables that concern the setting of the case are *Appointment Factors*, which contains *Number of Visits*, *Type of Visit*, and *Time between Appointments*; and *Documentation Status*, which contains *Documentation Age*, *Documentation Familiarity*, *Documentation Completeness*, and *Documentation Availability*. The documentation status variables are important in providing evidence about certain aspects of examinees' proficiencies in obtaining information. If we want to learn about an examinee's ability to seek and interpret information that can be obtained but is not initially present, then we cannot have *Documentation Completeness* set at "all information presented."

Task model variables that describe the patient include, as examples, *Age*, *Sex*, *Last Visit*, *Reason for Last Visit*, *Weight*, *Odors*, *Symptoms of Abuse/Neglect*, *Demeanor*, and *Risk for Medical Emergency*. Some of these are required for choosing stimulus material, such as photographs of the patient and responses to medical and personal

background questions. Others are important for focusing evidence. *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. These variables also play roles in determining the mix of cases to present to examinees. Every assessment might be required to include exactly one geriatric patient and one pediatric patient, for example, and the same number of male and female patients.

Characteristics of anticipated solutions are another important group of task-model variables. *Cues per Solution*, for example, indicates the complexity of a case that involves assessment or evaluation; it takes a value of "one" for simple cases, "few" for a moderately complex case, and "many" for a challenging case. Task-model variables like this help determine conditional probabilities in the Bayes net, either in informing expert judgments or as collateral information when estimating the probabilities from data. *Periodontal Assessment Procedures* contains a vector of procedures which can be carried out, and indicates which are "essential," "indicated," "irrelevant," and "contraindicated." This vector of values, determined by the task developer with expert advice, grounds evaluation rules for observables that concern treatment planning.

When a case uses a given task model, all the task-model variables associated with that task model must be assigned values. The list may be long, but setting the values need not be arduous. Some task-model variables are functions of lower level variables. *Case Complexity* is a higher level task-model variable that is useful for modeling task difficulty and for assembling tests; its value is derived from lower level task-model variables that indicate the specifics of the virtual patient's condition. Lower level task-model variables can be assigned default values based on typical distributions of normal variation, given basic conditions such as patient age and sex. Determining an underlying condition can imply constellations of expected values for many lower level task-model variables. These relationships could be automated in an authoring system that would allow the test developer to focus on exception conditions.

Besides task-model variables, task models include specifications for work products. The DISC simulator provides the sequence of actions an examinee takes, sometimes called a transaction list or an event trace. The task model describes its format and the code for its contents, including markers for phases of a case that can

be used to signal the need for appropriate evidence models. When an examinee responds to a case, a work product is produced that meets these specifications and contains the specifics of the examinee's actions. In turn, it will be parsed and evaluated by the evidence rules in the appropriate evidence models. Work-product specifications defined in the task model thus link the simulator at one end (the simulator must be capable of producing a product of this form) and evidence models at the other end (the parsing rules in an evidence model expect a work product with predefined format and kinds of content). They illustrate how carefully defining assessment design objects up front coordinates the work of very different kinds of experts, in this instance simulator designers and dental hygienists.

The CTA suggested the value of some additional, more structured work products. The DISC Scoring Team found consistent distinctions among novice, competent, and expert hygienists not only in the actions they chose, but the reasons they gave for choosing them. As seen earlier in Figure 2, many performance features concerned intermediate mental products, such as identification of cues, generation of hypotheses, and selection of tests to explore conjectures. Such steps often are not manifest in actual practice, but they directly involve the central knowledge and skills of problem solving in dental hygiene. In order to capture more direct evidence of this thought process than can be inferred from a transaction list alone, DISC will use work products that require the examinee to make normally mental steps explicit. Information-gathering actions during patient assessment and evaluation will need to be justified by specific hypotheses or as standard-of-care for the situation, and hypotheses will need to be justified by cues from available forms of information (in formats using nested lists of standard dental hygiene terms and procedures). Following patient assessment and evaluation, summary forms that require synthesizing findings will need to be completed (in a format similar to those of the insurance forms that are now integral to the practice of dental hygiene).

**The Simulator**

It may seem ironic that in a presentation about a simulation-based assessment, the shortest section is the one on the simulator itself. There are two reasons for this. The lesser is that other sources of information about the DISC simulator are already available, Johnson et al. (1998) chief among them. The more important reason is our desire to emphasize the evidentiary foundation that must be laid if we are to make sense of any complex assessment data. The central issues concern 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 possibilities, such as simulation-based scenarios, but these evidentiary considerations shape the thousands of decisions about how technologies can serve the purpose of the assessment.

In the case of DISC, the simulator needs to be able to create the task situations described in the task model. It also needs to capture that behavior in a form we have determined is necessary 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 in order to provide the data we need? As to the kinds of situations that will evoke the behavior we want to see, the simulator must be able to accomplish the following:

- 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 versus single visits so we can get evidence about examinees' capabilities to evaluate information over time; and
- Vary the virtual patient's state dynamically so we can learn about an examinee's ability to evaluate the outcomes of treatments that she chooses.

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

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

A key point is that DISC does not take the early version of the simulator as given and fixed. Ultimately, the simulator must be designed so that its highest priority is providing evidence about the targeted skills and knowledge—not

authenticity, not look and feel, not technology (Messick's 1994 discussion on designing performance assessments is mandatory reading in this regard).

## Putting the Pieces Together

The previous sections first reviewed a general schema for the evidentiary framework of complex assessments, then showed how the necessary building blocks are constructed around the substance and purpose of a particular domain and a particular product. This section shows how the pieces are fashioned and assembled in an operational assessment. This will be demonstrated by an overview of the DISC assessment design framework and scoring engine.

### Creating Tasks

The key to knowing how to score complex tasks is to design the tasks so you know they can evoke evidence about targeted knowledge and skills in ways you will be able to recognize and know how to accumulate. Figure 9 gives an overview of the process of creating tasks from this perspective, with numbers assigned to each step.

The first step in establishing a framework for task creation is fleshing out the objects of the CAF (1) in accordance with the substance of the domain and the purposes of the assessment. Task-model variables are used in any of the task models for describing features of tasks that are important for focusing evidence, determining difficulty, ensuring domain coverage, etc. (see Mislevy, Steinberg, & Almond, in press, on the roles of task-model variables). They are cataloged in the Case Feature Encyclopedia (2) as a common reference for task developers. The Proficiency/Evidence/Task Matrix (3) is simply a cross-referenced list derived from the student, evidence, and task models. A task developer who wants to write a task that taps into a particular aspect of proficiency can check this matrix to learn what observable variables are available to provide evidence about it, and which task models can be used to construct tasks that provide values for these observables.

DISC tasks are defined at a level compatible with distinguishable phases of interactions with patients; that is, initial patient assessment, treatment planning, treatment implementation, and follow-up evaluation. There can be more than one iteration of this cycle. A DISC case can therefore comprise more than one task. The following process is carried out for each. Having decided upon a certain task model through the reasoning described in the preceding paragraph (4), the developer instantiates the particulars that will make a unique task. This involves determining

values of the task-model variables that are involved in tasks written to this task model (5), and finding or creating suitable stimulus materials (6). The form of the work products will have been laid out in the work-product specifications of the chosen task model, but their specifics now need to be determined. The developer may need to work with domain experts at this point to determine the features of solutions, ideal and not so ideal (7), that will form the basis of case-specific evaluation rules (9). The DISC patient database (8) describes this case for the simulator. It contains the values of the task-model variables and the stimulus materials, specified as required to present and control the task in the DISC simulator environment, along with designation of the task model(s) and pointers to evidence models that will be needed to score performances.

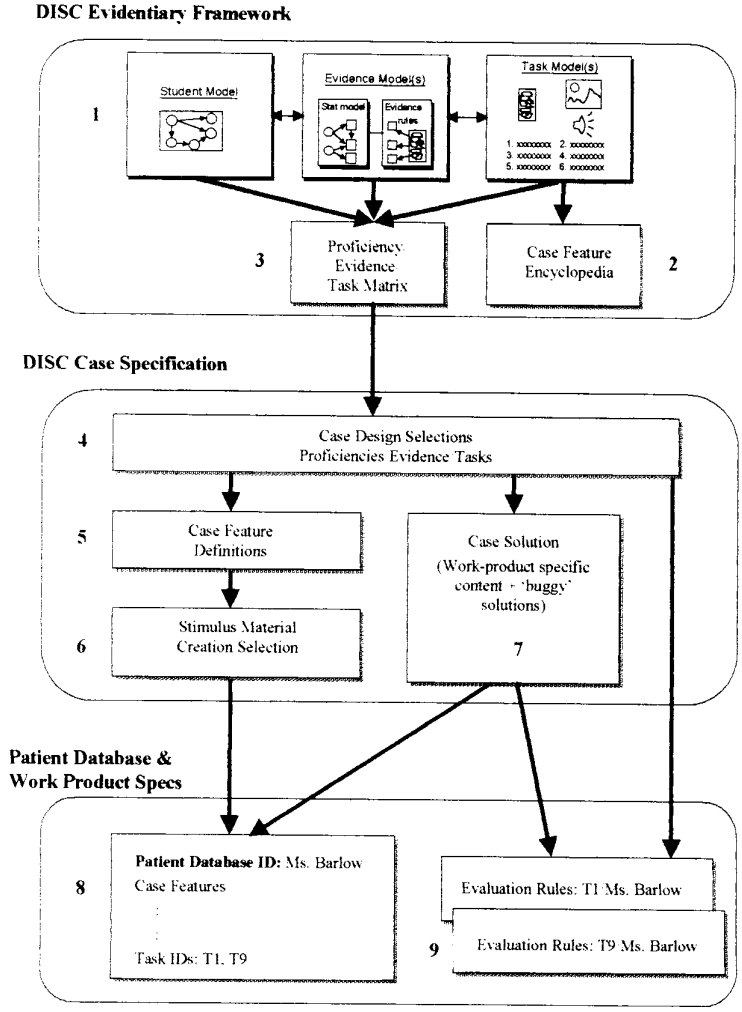



*Figure 9.* Overview of task creation.

## Calibrating Tasks

Cases are written within the frameworks that task and evidence models provide, as described earlier. The statistical submodel of the evidence model contains the structure of the appropriate Bayes net fragment, but not conditional probability distributions that are tailored to its particulars. That is, is it harder or easier than typical cases written with the same evidentiary skeleton, and does it provide a bit more or less evidence about the various student-model variables it informs? To some degree, these conditional probabilities can be based on expert knowledge and on previous empirical results from other tasks. More formally, we can estimate the conditional distributions from pretest data or field trials of the cases. Figure 10 depicts this process.

We consider estimating the conditional probabilities of a new case, say the "Ms. Barlow" case, when it is possible to collect data for both it and one or more previously calibrated case. The structure and the initial values for conditional probabilities are available for the student model (1) and the new case (2), the latter of which may contain more than one Bayes net fragment if it is a task that moves through multiple stages of interaction with the patient. Also available are conditional probabilities specific to the previously calibrated cases being presented (3), and field test data for all the cases from a sample of examinees (4). The conditional distribution of the new case can then be estimated (5); the reader is referred to Mislevy, Almond, Yan, and Steinberg (1999) for statistical details. The result is updated estimates of parameters that define the conditional probability distributions of the new task, along with updated estimates for the conditional probabilities in the student model and the previously calibrated tasks (6). The conditional probabilities for the new task are linked to it (7). With the particular values of its task-model variables, its case-specific evidence rules, and now case-specific conditional probabilities, the new task is ready to use for estimating the proficiencies of new examinees.

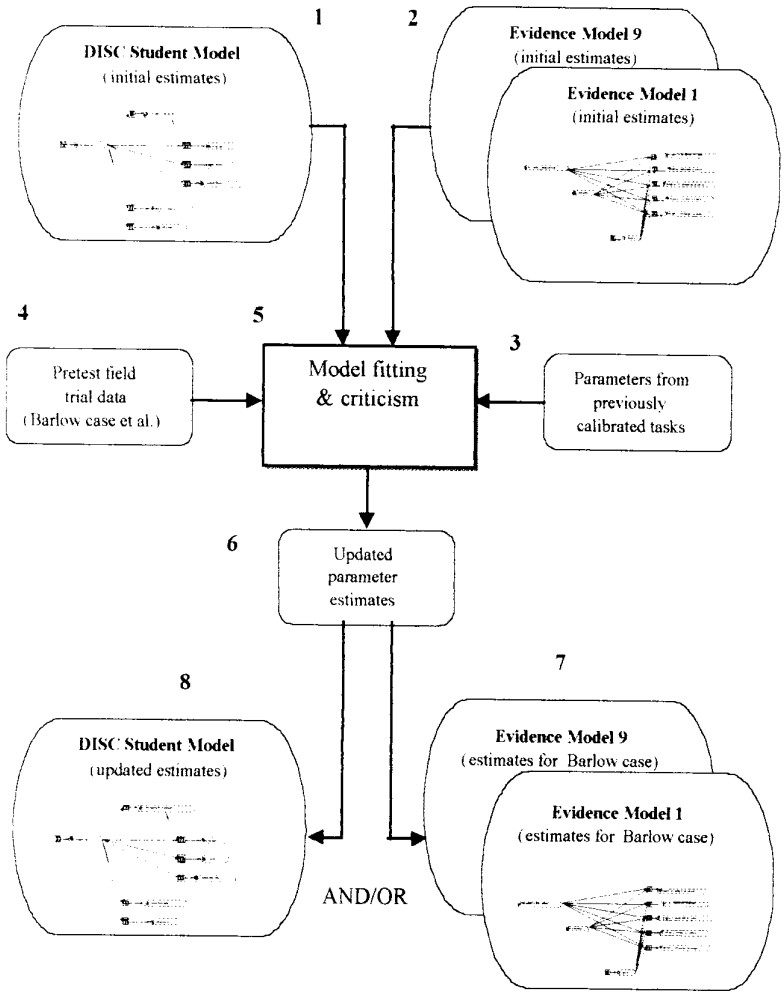



*Figure 10.* Calibrating a new case.

## Scoring an Examinee's Performance

The preceding sections have sketched the processes of task creation and calibration. They are prerequisites to assessing individual examinees; that is, presenting them with cases, gathering evidence about their knowledge and skills, and synthesizing this information in terms of the targeted proficiencies. Figure 11 gives an overview of the scoring process. The simulator references an algorithm that guides selection and sequencing of cases. A particular case that has been developed using the evidentiary framework is presented in the DISC simulation environment (1). The user's solution to a case is captured by the DISC simulator in the form of one or more task-specific work products (2). Scoring of performance on a case begins as the work products produced by the examinee through the DISC simulator are

examined for their evidentiary content (3). This is accomplished by processing each work product with task-specific rules of evidence. These rules of evidence evaluate a work product for the presence, absence, count, and/or quality of a pre-defined set of solution characteristics, or observable variables. This analysis produces a specific value for all observables associated with the task for which data are available in the work product.

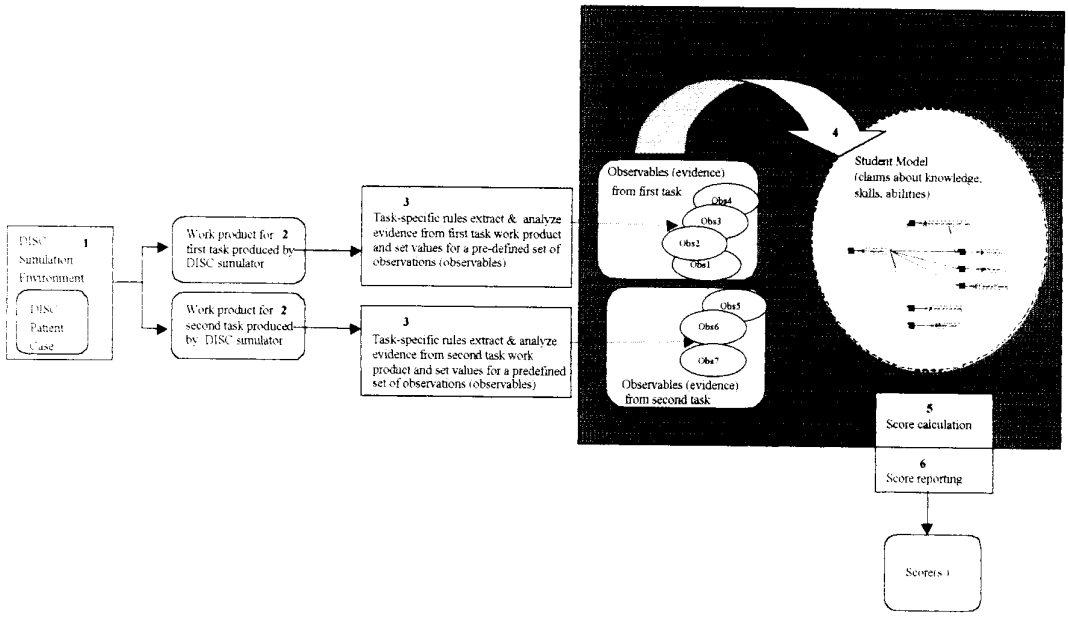



*Figure 11.* Overview of task scoring.

Once the body of evidence from a case is represented as realized values of observables, it is ready to be absorbed into the DISC Student Model. At this point, the DISC Student Model contains a probability distribution representing what is currently known about the user (4), possibly reflecting evidence that has been absorbed from previous cases. The DISC Scoring Engine *per se* (the shaded area of the diagram) consists of the DISC Student Model, pre-defined sets of observable variables with established relationships to Student Model variables (the structures of the evidence model Bayes net fragments), and evidence integration routines. The DISC Scoring Engine also provides for general-purpose functions that operate on the student model distribution (5) to calculate the values of customer-specified summary scores and measures of their precision (6).

## Next Steps

The preceding sections have described a rationale and framework for creating simulation-based problem-solving tasks in dental hygiene, evaluating examinees' actions, and synthesizing the results in terms of student-model variables. As this is written, the next steps in the development of the prototype are as follows.

The DISC simulator will be modified along the lines described above; that is, (a) to be able to present the range of situations and stimulus materials found necessary to evoke the targeted problem-solving knowledge and skills; (b) to enable the examinee actions that make manifest the use of that knowledge and skill; and (c) to produce the work products in the forms that capture the relevant actions and are amenable to evaluation in the planned ways.

Once the simulator has been modified, DISC will implement four initial cases that have been developed by the scoring team in accordance with the task and evidence models discussed above. The patient database for each of the virtual patients will be set up, the stimulus materials and presentation material will be created. The proto-rules for evaluating the observable variables in the appropriate evidence models will be specialized in their case-specific forms; they will be reviewed by experts for substantive accuracy, and tested against sample work products for adequacy. Small groups of examinees will be tested in order to refine cases, rules, and interfaces as necessary.

Once the cases are ready, a "contrasting groups" comparison study will be carried out. About 10 hygienists each at the novice, recently licensed, and expert levels of proficiency will work through each of the problems. Their responses will be used to refine the conditional probabilities in the Bayes nets of the evidence models.

From this study, DISC can begin to explore some issues that would be critical in an operational licensing assessment. First, various summary functions of the student-model variables can be defined and resulting scores for the participants can be obtained. DISC will look for discrimination between the novices, who presumably are not yet ready to be licensed, and recently licensed and expert hygienists, who presumably are. Which aspects of proficiency seem to be most useful in distinguishing presumed failures from presumed passers? What projections capture the distinctions? Do the empirically effective distinctions accord with those that are thought to be important from a substantive point of view?

Second, having estimated conditional probabilities in a variety of evidence models and chosen some plausible summary projections, DISC can carry out simulation experiments to study how many tasks of which kinds provide what levels of accuracy for the pass/fail decision. The conditional probabilities estimated for the evidence models in the study can be replicated for any number of hypothetical tasks, and simulated responses can be generated for any number of hypothetical examinees with various profiles of proficiency. From such studies DISC can calculate projected rates of consistency and accuracy in pass-fail decisions.

Reporting formats for both final scores and intermediate-level feedback can then be designed for potential use in an operational assessment. If operational use is desired, then DISC will determine the procedures, capabilities, and structures that would be required for scaling up. The requirements analysis could begin in parallel with the above described work, or follow it after a favorable go/no-go decision.

## Conclusion

What is the payoff we hope to gain from all of this work? Basically, a framework for creating an indefinite series of unique, realistic cases, each complex and interactive, and each having a predetermined method of scoring. The skeleton of the evidentiary argument, and a way to incorporate particulars, already has been laid out. We will have produced a reusable student-model, which we can use to project an overall score for licensing, but which supports mid-level feedback as well. We will have produced reusable evidence and task models around which DISC can write indefinitely many unique cases. We'll also have produced a framework for writing case-specific evaluation rules. The technology for scoring, work-product evaluation, and simulation can be applied in other products and in other learning domains.

We may conclude by contrasting two approaches for making sense of complex assessment data in ongoing, large-scale applications. The hard way is to ask, "How do you score it?" after you've built the assessment and scripted the tasks or scenarios. Unfortunately, the contrasting approach isn't the easy way. It's a different hard way: Design the assessment and the tasks or scenarios around what you want to make inferences about, what you need to see to ground those inferences, and the structure of the interrelationships. This still isn't easy, but it just might work.

## References


Almond, R.G., Herskovits, E., Mislevy, R.J., & Steinberg, L.S. (1999). Transfer of information between system and evidence models. In D. Heckerman & J. Whittaker (Eds.), *Artificial Intelligence and Statistics 99* (pp. 181-186). San Francisco: Morgan Kaufmann.

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

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

Edwards, W. (1998). Hailfinder: Tools for and experiences with Bayesian normative modeling. *American Psychologist, 53,* 416-428.

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.

Gardner, H. (1991). *The unschooled mind: How children think, and how schools should teach.* New York: Basic Books.

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. (1997). Cognition and learning. In D. Berliner & R. Calfee (Eds.), Handbook of educational psychology (pp. 15-47). New York: Simon & Schuster Macmillan.

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

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

Kadane, J.B., & Schum, D.A. (1996). *A probabilistic analysis of the Sacco and Vanzetti evidence.* New York: Wiley.

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases.* Cambridge: Cambridge University Press.

Lesgold, A.M., Rubinson, H., Feltovich, P.J., Glaser, R., Klopfer, D., & Wang, Y. (1988). Expertise in a complex skill: Diagnosing X-ray pictures. In M.T.H. Chi, R. Glaser, & M.J. Farr (Eds.), *The nature of expertise* (pp. 311-342). Hillsdale, NJ: Erlbaum.

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

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

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

Mislevy, R.J. (1995). Probability-based inference in cognitive diagnosis. In P. Nichols, S. Chipman, & R. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 43-71). Hillsdale, NJ: Erlbaum.

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., 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., & Almond, R.G. (in press). On the several roles of task model variables in assessment design. In S. Irvine & P. Kyllonen (Eds.), *Generating items for cognitive tests: Theory and practice.* Hillsdale, NJ: Erlbaum.

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

Pennington, N., & Hastie, R. (1991). A cognitive theory of juror decision making: The story model. *Cardozo Law Review, 13*, 519-557.

Schum, D.A. (1987). *Evidence and inference for the intelligence analyst.* Lanham, Md.: University Press of America.

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

Spiegelhalter, D.J., Thomas, A., Best, N.G., & Gilks, W.R. (1995). *BUGS: Bayesian inference using Gibbs sampling, Version 0.50.* Cambridge: MRC Biostatistics Unit.

## Appendix A

### A Simple Example Using Bayes Nets in Assessment

This appendix offers a simple example of how Bayes nets can be used in assessment. The interested reader is referred to Mislevy (1994, 1995) and Mislevy and Gitomer (1996) for additional discussion and examples. There is just one student-model variable in this small example, "level of proficiency." It has two levels, expert and novice, and we assume a student is unambiguously one or the other. The work product is the examinee's sequence of actions in taking a patient history in a particular task situation, and we assume we can determine unambiguously whether a given sequence is adequate or inadequate. What we want to do in the assessment setting is observe an examinee's history-taking, evaluate its adequacy, and update our belief about the examinee's expert/novice status. Following is a numerical illustration of how one moves from state of relative ignorance about the unknown value of the student model variable to a state of greater knowledge by incorporating value of evidence from values of the observed variable.

Figure A1 is a matrix of conditional probabilities for taking an adequate or inadequate patient history in a particular situation of interest, given that the actor is an expert or is a novice. The top row gives conditional probabilities of .8 and .2 for observing an adequate and inadequate history, respectively, taken by a subject known to be an expert. We assume for the moment that we know these values cold; we've just run an experiment in which we've observed 10,000 acknowledged experts taking patient histories and noted that 8,000 of them took adequate histories and 2,000 took inadequate ones. Similarly, the bottom row gives conditional probabilities of .4 and .6 for observing an adequate and inadequate history, respectively, to be taken by a novice. Note that this is reasoning from proficiency to expectations for observables, just the opposite of what we want to do in assessment.

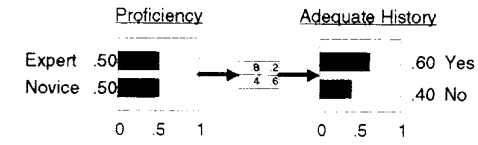
| | Adequate History | |
|---|---|---|
| Proficiency | Yes | No |
| Expert | .8 | .2 |
| Novice | .4 | .6 |

Figure A1. Conditional probability matrix.

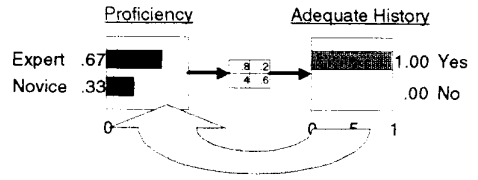
Now let's reason in the other direction. We've done our experiment, so we know experts take adequate histories 80% of the time and novices do 40% of the time. If a new examinee with an unknown proficiency takes an adequate history, how should this evidence influence our belief about his or her level of expertise? The "adequate history = yes" column gives the answer: we should shift our beliefs by a factor of .8/.4, or 2-to-1, toward the expert category. Technically, this column is the likelihood vector that observing an adequate history induces. Analogously, if we observe that an inadequate history has been taken, we should modify our beliefs by a factor of .2/.6, or 1-to-3, shifting in the direction of novice.

The first panel of Figure A2 is a representation of the probability distributions that contain our beliefs following the conditional probabilities experiment but before observing the response of the next new examinee. It assumes 50-50 chances that this new examinee is an expert or a novice, as indicated by the two probability bars at .5 for the two possible proficiency values. In the event that the examinee is an expert, she will take an adequate history with probability .8. In the event that she is a novice, she will take an adequate history with probability .4. Together this set of beliefs implies a .6 expected probability of observing an adequate history. This is what the probability bars in the distribution for the observable variable indicate.

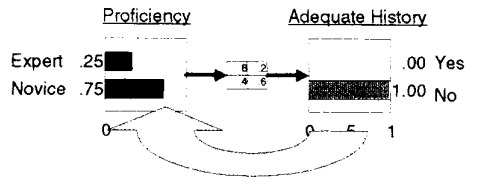
The second panel of Figure A2 shows how our beliefs change if we observe an examinee take an adequate history. The probability for the possible values of adequate history is all on "yes" now, because we've actually observed it. Our belief that the examinee is an expert has shifted up from 50-50 accordingly, using Bayes theorem to shift the probabilities for expert/novice in the 2-to-1 ratio. The third panel shows how our beliefs change if instead we observe an examinee take an inadequate history. Belief shifts downward, in proportion to the 1-to-3 ratio of conditional probabilities of an inadequate history from an expert and from a novice respectively.




A. Before Observing Response




B. After Observing 'Yes' Response




C. After Observing 'No' Response

Figure A2. Updating beliefs.

In applied work, for a specific inferential problem, the structure of the Bayes net and the conditional probabilities can often be taken as known. Where do the conditional probabilities come from? Initial estimates of conditional probabilities can come from expert opinion, and they can be refined with pretest data (Mislevy, Almond, Yan, & Steinberg, 1999). This is analogous to estimating item parameters in IRT. It can be accomplished via Monte Carlo Markov Chain estimation, using, for example, the BUGS computer program (Spiegelhalter, Thomas, Best, & Gilks, 1995).

## Appendix B

DISC Evidence Models

**Information gathering...**

*In Assessment* (1)

With Ethics/Legal issue (2)

With Medical issue (3)

With Ethics/Legal + Medical issues (4)

*In Evaluation* (5)

With Ethics/Legal issue (6)

With Medical issue (7)

With Ethics/Legal + Medical issues (8)

**Hypothesis generation...**

*In Assessment* (9)

With Ethics/Legal issue (10)

With Medical issue (11)

With Ethics/Legal + Medical issues (12)

*In Evaluation* (13)

With Ethics/Legal issue (14)

With Medical issue (15)

With Ethics/Legal + Medical issues (16)

*Data recording* ***(17)***

With Ethics/Legal issue (18)

*Information recognition* ***(19)***

In Assessment (20)

In Evaluation (21)

**Hypothesis testing...**

*In Assessment* (22)

With Ethics/Legal issue (23)

With Medical issue (24)

With Ethics/Legal + Medical issues (25)

*In Evaluation* (26)

With Ethics/Legal issue (27)

With Medical issue (28)

With Ethics/Legal + Medical issues (29)

*Treatment implementation* (30)

*Assessment summary form* (31)

*Treatment planning* (32)

With Ethics/Legal issue (33)