A Cognitive Task Analysis, With Implications for Designing A Simulation-Based Performance Assessment

CSE Technical Report 487

Robert J. Mislevy and Linda S. Steinberg
Educational Testing Service, Princeton, New Jersey

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

Russell G. Almond
Educational Testing Service, Princeton, New Jersey

Lynn Johnson
Dental Interactive Simulations Corporation, Aurora, Colorado

August 1998

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
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 work reported herein was supported under 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.

A COGNITIVE TASK ANALYSIS, WITH IMPLICATIONS FOR DESIGNING A SIMULATION-BASED PERFORMANCE ASSESSMENT[1]

Robert J. Mislevy and Linda S. Steinberg
CRESST/Educational Testing Service, Princeton, New Jersey

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

Russell G. Almond
Educational Testing Service, Princeton, New Jersey

Lynn Johnson
Dental Interactive Simulations Corporation, Aurora, Colorado

ABSTRACT

To function effectively as a learning environment, a simulation system must present learners with situations in which they use relevant knowledge, skills, and abilities. To function effectively as an assessment, such a system must additionally be able to evoke and interpret observable evidence about targeted knowledge in a manner that is principled, defensible, and fitting to the purpose at hand (e.g., licensure, achievement testing, coached practice). This article concerns an evidence-centered approach to designing a computer-based performance assessment of problem-solving. The application is a prototype licensure test, with supplementary feedback, for prospective use in the field of dental hygiene. We describe a cognitive task analysis designed to (a) tap the knowledge hygienists use when they assess patients, plan treatments, and monitor progress, and (b) elicit behaviors that manifest this knowledge. After summarizing the results of the analysis, we discuss implications for designing student models, evidentiary structures, task frameworks, and simulation capabilities required for the proposed assessment.

[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, made up of Barry Wohlgemuth, D.D.S., DISC President and Project Director; Lynn Johnson, Ph.D., Project Manager; Gene Kramer, Ph.D.; and five core dental hygienist members: Phyllis Beemsterboer, R.D.H., Ed.D., Cheryl Cameron, R.D.H., Ph.D., J.D., Ann Eshenaur, R.D.H., Ph.D., Karen Fulton, R.D.H., B.S., and Lynn Ray, R.D.H., B.S.. The team was expanded for this work with three additional experts in dental hygiene: Susan Callahan Barnard, R.D.H., M.S., Hope-Clair Holbeck, R.D.H., M.S., and Dawn Walters, R.D.H., M.Ed. We would like to express our gratitude to the thirty-one dental hygienists who volunteered to work through the cases of the cognitive task analysis. We also thank Denny Way, Paul Jones, and David Williamson of The Chauncey Group International and Nancy Slach of the University of Iowa for their contributions.

Don Melnick (1996), who for many years directed the National Board of Medical Examiners' computer-based patient management assessment project, recently remarked, "It is amazing to me how many complex "testing" simulation systems have been developed in the last decade, each without a scoring system" (p. 117). A foundation for sound assessment must be laid long before a simulator is complete and tasks are written. This article summarizes research designed to lay the foundation for a computer-based performance assessment of problem-solving. A quote from Messick (1992) captures the essence of the *evidence-centered* approach to assessment design that guides our work:

> [We] would begin by asking what complex of knowledge, skills, or other attributes should be assessed, presumably because they are tied to explicit or implicit objectives of instruction or are otherwise valued by society. Next, what behaviors or performances should reveal those constructs, and what tasks or situations should elicit those behaviors? Thus, the nature of the construct guides the selection or construction of relevant tasks as well as the rational development of construct-based scoring criteria and rubrics. (p. 17)

We focus on the methods, results, and implications of a cognitive task analysis that addresses these evidentiary issues, to ground the design of a simulation-based assessment of problem-solving and decision-making in dental hygiene.

BACKGROUND

In 1990, a consortium of dental education, licensure, and professional organizations created the Dental Interactive Simulation Corporation (DISC) to develop computerized assessments and continuing education products that simulate the work dentists and dental hygienists perform in practice. 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 oral cavity; and performing prophylaxes, scaling, and root planing. Currently, multiple-choice examinations probe hygienists' content knowledge required by 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 article is written, DISC has developed a prototype of a dental simulation system in the context of continuing education (Johnson et al., in press). 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. These capabilities provide a starting point for the proposed simulation-based dental hygiene licensure assessment.

A simulation-based assessment must elicit behavior that bears evidence about key skills and knowledge, and it must additionally provide principled interpretations of that evidence in terms that suit the purpose of the assessment. Figure 1 sketches the basic structures of an evidence-centered approach to assessment design (Almond, Mislevy, & Steinberg, 1998). Designing these variables and models and their interrelationships is a way to answer Messick's questions, and to build coherent assessments around the answers:

- "What complex of knowledge, skills, or other attributes should be assessed?"

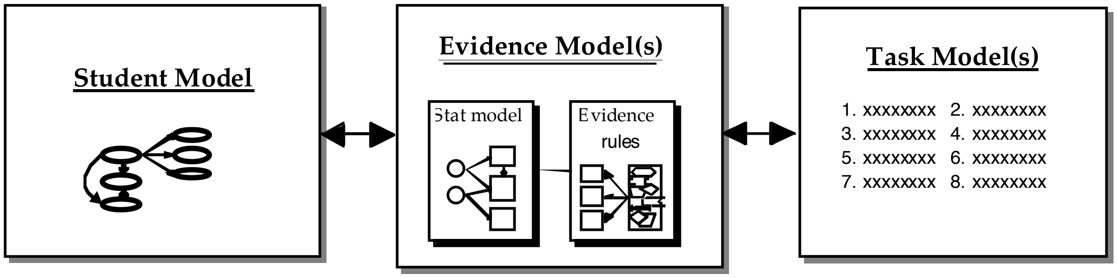
  A given assessment is meant to support inferences for some purpose, such as a licensing decision, diagnostic feedback, guidance for further instruction, or some combination. Student-model variables describe characteristics of examinees—knowledge, skills, and abilities, which we will call *knowledge* collectively for short—upon which these inferences are to be based. The student model expresses the assessor's knowledge about an examinee's values on these variables.

- "What behaviors or performances should reveal those constructs?"

  An evidence model expresses how what is observed in a given task constitutes evidence about student-model variables. Observable variables describe features of specific task performances.

- "What tasks or situations should elicit those behaviors?"

  Task-model variables describe features of situations that will be used to elicit performance. A task model provides a framework for characterizing and for constructing situations with which a candidate will interact to provide evidence about targeted aspects of knowledge.



Student model variables describe characteristics of examinees—e.g., knowledge, skills, abilities—that one wants to make inference about (reporting, diagnostic statements, decisions, etc.)

Values of observable variables are modeled as probabilistic functions of student model variables.

Observable variables describe features of examinees' performances. An evidence model expresses how what is observed in a given task constitutes evidence about student model variables.

Evidence rules extract salient features from an examinee's work product, and evaluate values of observable variables.

The statistical model expresses the probabilistic dependence of observable variables on student model variables.

Task model variables describe features of tasks. A task model provides a framework for characterizing constructing the situations in which examinees will act.

Task models (a) guide task construction, (b) focus the evidentiary value of tasks, (c) support test assembly specifications, and (d) mediate the relationship between tasks and student model variables.

Figure 1. Models and variables in evidence-centered assessment.

The assessor uses task models to design situations that have the potential to provide evidence about an examinee's knowledge. She uses evidence models to identify, from what the examinee does in these situations, evidence about targeted aspects of that knowledge. She uses the student model to accumulate and represent belief about those aspects of knowledge, expressed as probability distributions for student-model variables (Almond & Mislevy, in press; Mislevy & Gitomer, 1996).

Overview of Procedures

Given a purpose and intended use of an assessment, an understanding of the evidentiary requirements for targeted aspects of knowledge grounds the rationale for designing tasks and the environment in which they are performed. Previous research, accumulated experience, and empirical study can support this understanding (e.g., O'Neil, Allred, & Dennis, 1997). We designed a cognitive task analysis to flesh out the general assessment structures described above with the specifics of problem-solving and decision-making in dental hygiene, for the primary purpose of licensure with supplementary feedback. 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 use to carry out those tasks. A cognitive task analysis 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 hard (Newell & Simon, 1972).

With creating assessment structures as the ultimate objective, we adapted cognitive task analysis methods from the expertise literature (Ericcson & 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. Five phases of work are described in the following sections.

Phase 1. Outlining relevant areas of knowledge and features of situations that can evoke evidence about them, to guide the design of the cognitive task analysis.

Phase 2. Creating a set of cases for the cognitive task analysis, tailored to address the evidentiary issues in assessment design.

Phase 3. Gathering talk-aloud protocols from expert, competent, and novice hygienists working through the cases.

Phase 4. Analyzing the protocols to characterize patterns of problem-solving behavior that distinguish hygienists at different levels.

Phase 5. From the results of the relatively unconstrained cognitive task analysis, drawing implications for the more structured context of high-stakes, large-scale, assessment.

PHASE 1: PLANNING THE COGNITIVE TASK ANALYSIS

A group of dental hygiene experts assembled by DISC—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. They produced a specifications matrix with *actions* as its rows and *signs* as its columns (Figure 2). The actions are broad categories of hygienists' interactions with patients (Darby & Walsh, 1995, p. 50), and are further detailed in Figure 3. The signs are categories of standard ways that knowledge is captured, represented, expressed, or conveyed in the field. This matrix organization was meant to aid thinking about the kinds of knowledge representations that are used in various phases of interactions with patients. Specifications such as these are a typical element of a job analysis, laying out the work that people in a profession do and the circumstances in which they do it (Schneider & Konz, 1989). Using the *actions* dimension of the specifications map, the Team discussed examples of behaviors that tend to distinguish among three levels of proficiency: Expert, Competent (recently licensed practitioners), and Novice (students still in the course of study). The team also delineated contextual and constraining factors that can affect all phases of patient/hygienist interactions, including setting and time constraints, and the patient's periodontal status, age, medical condition, anxiety, economic status, and cultural factors.

	Signs					
Actions	Patient medical and dental history	Comprehensive oral examination findings	Radiographs	Oral photos	Lab reports	Notes and documentation
Patient assessment						
Data interpretation						
Planning hygiene care						
Implementation						
Evaluation*						
Communication						

Note. Patient Assessment concerns ascertaining patient status *before treatment*, while Evaluation concerns ascertaining patient status *during and after treatment*, in light of expectations for that treatment.

Figure 2. Content specification map for clinical dental hygiene settings.

Assessment (ascertaining patient status *before treatment*)

- A. Medical, dental, and social history
- B. Extraoral examination
- C. Intraoral examination
- D. Dentition examination
- E. Periodontal evaluation (including deposits and stains)
- F. Occlusal evaluation
- G. Clinical testing (e.g., thermal, vitalometer, percussion)
- H. Radiographs

Data Interpretation

- A. Interpreting case documentation
- B. Recognizing normalities and abnormalities
- C. Evaluating quality of case documentation

Planning Hygiene Care

- A. Planning individualized instruction
- B. Categorizing and prioritizing dental hygiene treatment

Implementation

- A. Infection control
- B. Treatments
  - 1. Dental hygiene services
  - 2. Support services
- C. Education and prevention

  Instruction for prevention and management of oral/systemic diseases/conditions

Evaluation (ascertaining patient status *during and after treatment*, in light of expectations)

- A. Short-term, long-term goals
- B. Comparative patient status in response to intervention
- C. Comparative patient condition over time

Communication

- A. With client
- B. With other professionals

Figure 3. Breakdown of "Actions" in dental hygiene content specification map.

PHASE 2: CREATING THE CASES FOR THE COGNITIVE TASK ANALYSIS

Working from these resources, the 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. A case is defined in terms of features of not only stimulus materials, but of context, expectations, instructions, affordances, and constraints the examinee will encounter. In general, a task feature is any characteristic of the task situation that prompts an examinee's performance or describes the environment in which the performance is carried out.

A guiding principle of evidence-centered design, whether for a licensure exam, a tutoring system, or a cognitive task analysis, is that the characteristics of the required evidence determine the characteristics of tasks. It is all too easy to obtain agreement on general statements of valued knowledge or to describe tasks without explicating how performances in those tasks constitute evidence about targeted aspects of knowledge. A focus on evidence is especially important for designing complex open-ended tasks because so much information is necessary to explicate a task. The more unconstrained and complex the task, the more rigorous the design criteria must be to make sense of the examinee's actions. The designer must understand how each feature defined for a task contributes to the task's purpose, eliciting performances that possess the right evidentiary characteristics to distinguish among the kinds or levels of proficiencies of interest.

The Scoring Team built the cognitive task analysis cases around features that would provoke decision-making and stress the areas of knowledge identified as important in the domain. Some useful categories of task features are setting, substance, tools and operations, interaction, range of responses, and help/guidance/ structuring. Table 1 defines these categories and notes their roles in the analysis. Task features for setting, substance, and tools and operations will have close counterparts in a simulation-based licensure exam, since they concern the substantive nature of the decision to be made or the problem to be solved. Task features dealing with interaction, range of responses, and help/guidance

Table 1

CTA Task Features and Implications for Task Model Variables

Category	Description	Features of CTA tasks	Imputed task model variables
Setting	The context in which the work takes place	The offices of either a dentist or periodontist in private practice.	Primarily, the location of action. The CTA tasks suggest the main setting is an operatory in a private dental or periodontal office. Other values might be clinics and nursing homes.
Substance	Characteristics of the situation or problem stimuli the examinee must deal with	Characteristics of a specific virtual patient, including medical, dental, and periodontal status, social/occupational background, and appointment- or treatment-related factors and attitudes.	CTA tasks suggest subcategories of medical condition, general symptoms, oral condition, dental condition, periodontal condition, background factors, medication, documentation, appointment factors, and chief complaint. Each may be further decomposed before values are assigned. Medication may be broken down into type (blood pressure, analgesic, birth control) before each type is assigned values (e.g., aspirin, Tylenol, and ibuprofen for analgesic).
Tools and operations	Resources and actions available to the examinee for responding to the stimulus	Documentation, information, equipment, and procedures available to the hygienist. Focus on defining appropriate documentation (medical histories, radiographs, lab reports, etc.), procedures for acquiring and referencing such information, and dental hygiene procedures that might be carried out with patients.	Key requirements concern sources of information for patient assessment, procedures for acquiring and using it, and capabilities to record and transmit it. Potential variables include sources of information (reference materials, patient, internal professional staff); reports and forms (radiographs, dental charts, lab reports, referrals); search and retrieval procedures (interview, written request, observation, diagnostic procedure); recording information (specialized notation, patient notes); transmitting information (verbal, written, demonstration); connecting information (data elements with assessment, assessment elements with treatment plan).

(table continues)

Table 1 (continued)

Category	Description	Features of CTA tasks	Imputed task model variables
Interaction	Extent and mechanism for the examinee to interact with the task situation	Individuals, simulated or real, with whom the subject can or should interact during task performance: the patient, persons related to the patient, professional staff internal and/or external to the setting. The main interaction was the expert interviewer providing information as the case progressed: about patient data, responses to treatments the subject ordered, answers to questions the subject asked, information coming from referrals, reactions the patient would have to actions the subject undertook, etc.	This is the extent to which interaction is required in task performance (e.g., information given, equipment used, referrals & resources available). Sample values might include: none, minimal, moderate, high. More than one interaction variable may be defined if different degrees of interaction are required at different points in the case, or with respect to different sources of information.
Range of responses	The range and nature of the elements of examinee performance that are captured	Verbal protocols; identification of the data the subject requested and examined as to sequence and point in the case. No actual assessment or treatment procedures carried out.	Since all performances will be captured in a transaction list, these variables will be common across tasks. Evidence rules that extract and evaluate actions from the work product must be developed in concert with specific tasks to address which actions contained in the work product are relevant.
Help/ guidance/ structuring	(1) Instructions and directive feedback to the examinee before or during task performance (2) The ways in which tools and operations enable, facilitate, or channel examinee actions	After the opening scenario for a case was presented, the subject's sequence, timing, and substantive focus were unconstrained. The researcher prompted subjects for explanations and rationales of their actions.	These variables indicate the pacing and the nature and degree of structuring of a case; e.g., minimally constrained; stepped through distinct phases with summaries for each; patient assessment only; treatment planning from given assessment information.

had to be worked out as procedures for conducting the cognitive task analysis. In Phase 5 we will discuss how analogous issues arise in a simulation-based licensure test, but have different solutions because cognitive task analysis and large-scale licensure testing have different purposes, resources, and constraints.

By focusing on the type of evidence a case was meant to elicit, the scoring team members determined appropriate values and roles for selected task features. Some features were incidental to a case, adding detail and realism without being central to the core issue; others were used to set the complexity of a case or to evoke evidence about targeted aspects of knowledge. As a first example of a feature intended to provoke evidence about particular aspects of dental hygiene knowledge, Case #5 included evidence of bruxism (grinding teeth) in intraoral photographs. Figure 4 shows the characteristic worn surfaces. This case thus provides the opportunity to observe whether a subject detects the condition, explores connections with patient history and lifestyle, and discusses implications with the patient. As a second example, the stimulus materials of Case #2 indicate a rapid deterioration of periodontal status over a six-month period, a consequence of uncontrolled diabetes. This information is implicit in significant changes in pocket depths from previous to current periodontal charts. Neither the charts, nor radiographs and oral photographs that provide additional clues, are presented unless the subject specifically requests them or procedures that will produce them. This collection of task features provides an opportunity

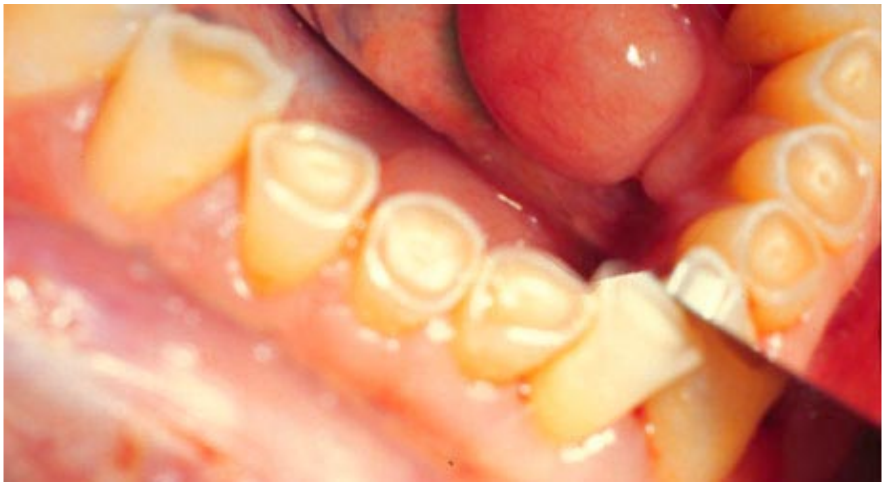


Figure 4. Intraoral photograph from Case #5, showing signs of bruxism (grinding teeth).

to observe data-gathering procedures, and whether the subject integrates information over time and across sources, and relates the patient's oral status to a medical condition.

PHASE 3: CAPTURING SOLUTIONS

Five Scoring Team members served as expert interviewers. They solicited a total of 31 subjects, all of whom were women, to interview. Eleven were classified as expert, nine as competent, and 11 as novice. Experts typically taught in dental hygiene educational programs or practiced with the Team member who recruited them. Competent subjects were recently licensed hygienists. Novices were dental hygiene students who had not yet completed their education.

A subject, an expert Scoring Team member, and one or two psychologist researchers participated in each interview. All interviews were audiotaped and later transcribed. The expert interviewer began by introducing the subject to the researcher, and asking her to complete consent and background information forms. The researcher explained that the subject would be seeing nine cases and was to talk out loud as she worked through them. A typical session took about two hours.

The expert dental hygienist provided the brief prepared verbal description of the patient in each case in turn. The researcher asked the subject to describe her thoughts out loud, and say 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 expert interviewer 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. Expertise in dental hygiene was essential to present cases in this manner, since the expert interviewer had to provide feedback that was both plausible and substantively correct along any path a subject might take. The researcher would ask the subject to interpret the information; for example, to say what she thought 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.

PHASE 4: ANALYZING THE PROTOCOLS

Procedures

Audiotapes and transcripts in hand, the Scoring Team and the researchers met for three days to analyze the protocols. The mission was to abstract, from specific actions of individual subjects in particular cases, general characterizations of patterns of behavior, 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 committee sought patterns that would be useful in distinguishing hygienists at different levels of competence. We refer to the resulting characterizations as *performance features*.

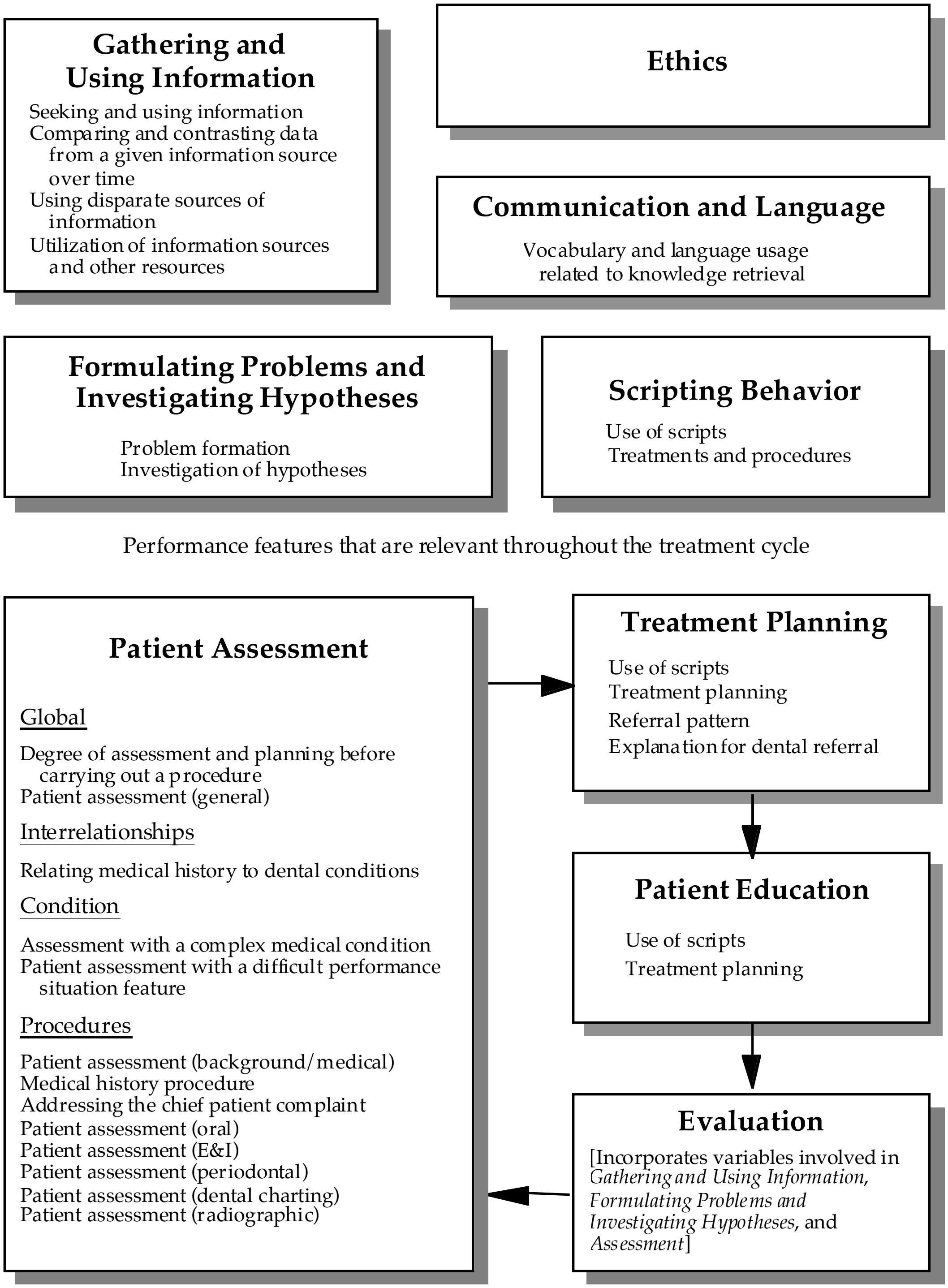
To define performance features, the Scoring Team and researchers began by breaking into two groups of four people each. On the first day, one group reviewed protocols from one case and the other group reviewed protocols from another case. Each group listened to tapes or read transcriptions aloud, starting with interviews from two random subjects from each level of proficiency. After listening to the interviews and taking notes, the groups discussed the protocols, abstracting and illustrating performance features that differentiated expert, competent, and novice hygienists in that case. The groups studied additional transcripts of subjects working through the same cases, eventually addressing about 20 of the 31 available. The groups came together to review the performance features they had defined separately, and combine them into a single document. Over the next two days the groups analyzed the remaining seven cases. They considered whether the behaviors they saw in new cases (a) revealed additional distinctions among expert, competent, and novice hygienists, (b) added examples of previously-defined performance features, or (c) suggested refinement or extension of previously-defined features. The Team also identified knowledge areas underlying these performance features, to aid in defining student model variables.

Summary of Findings

The following discussion of the performance features is organized in terms of the Scoring Team's categorization shown in Figure 5. They summarize the results of Phase 4 of the project. (Phase 5 will provide further refinement and classification in terms of implications for observable variables.) We will note that these results on problem-solving behaviors and knowledge requirements in dental hygiene reflect patterns that have surfaced in studies of expertise across many domains; that is,

> In brief, [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 Scoring Team defined the category **Gathering and Using Information** to consist of four performance features, of which *seeking and using information* is most broadly construed. The Team's summary of expert behavior includes the phrase "seeks relevant information," where "relevant" is seen from an expert's perspective. These performance features indicate the degree to which a hygienist possesses knowledge structures that can incorporate available data and suggest syndromes or etiologies to apprehend a situation. These structures in turn suggest correlates that guide further pursuit of information. The three more specific performance features concern pursuing and integrating information from a given source across time points, across disparate sources, and from sources in addition to standard dental hygiene data.


Gathering and Using Information
Seeking and using information
Comparing and contrasting data from a given information source over time
Using disparate sources of information
Utilization of information sources and other resources
Ethics
Communication and Language
Vocabulary and language usage related to knowledge retrieval
Formulating Problems and Investigating Hypotheses
Problem formation
Investigation of hypotheses
Scripting Behavior
Use of scripts
Treatments and procedures
Patient Assessment
Global
Degree of assessment and planning before carrying out a procedure
Patient assessment (general)
Interrelationships
Relating medical history to dental conditions
Condition
Assessment with a complex medical condition
Patient assessment with a difficult performance situation feature
Procedures
Patient assessment (background/medical)
Medical history procedure
Addressing the chief patient complaint
Patient assessment (oral)
Patient assessment (E&I)
Patient assessment (periodontal)
Patient assessment (dental charting)
Patient assessment (radiographic)
Treatment Planning
Use of scripts
Treatment planning
Referral pattern
Explanation for dental referral
Patient Education
Use of scripts
Treatment planning
Evaluation
[Incorporates variables involved in Gathering and Using Information, Formulating Problems and Investigating Hypotheses, and Assessment]

Performance features that are relevant throughout the treatment cycle

Performance features that apply to particular phases of the treatment cycle

Figure 5. Performance features in decision-making and problem-solving in dental hygiene.

Three examples from Case #2 show how hygienists at different levels of expertise dealt with rapid deterioration of periodontal status that is implicit in information across time points and across disparate sources. Below, Expert A inspects radiographs and periodontal charts, and examines the patient visually—three sources of information. Not finding the coordination she expects (due to the atypical six-month difference) she goes back to do a full periodontal charting to verify the previous findings.

Expert A: At this point warning bells go off ... looking at these x-rays, have been seeing this person every 6 months and have this extensive bone loss.

Interviewer 2: Right so, so, there are no x-rays in the file.

Expert A: There is no x-rays in the file, the patient was seen in this office 6 months ago, this is the charting, this is the perio charting from 6 months ago?

Interviewer 1: Correct.

Expert A: So this perio charting was done 6 months ago, prior to today.

Interviewer 2: Yeah, you're looking now at the perio chart.

Expert A: So I would go back, I would take a look at her mouth and something doesn't look right here. Now I would take x-rays.

Interviewer 2: Okay, so now you take x-rays.

Interviewer 1: These are the x-rays.

Expert A: Looking at these x-rays and looking at this 6-month perio charting, I would be quite concerned, because this doesn't match these x-rays.

Interviewer 2: So what would you do?

Expert A: I would do another full perio charting.

The new periodontic evaluation confirmed the bone loss, and Expert A went on to explore the possibility of systemic medical problems to explain the change in oral status.

Next, competent hygienist B compares radiographs to the periodontal charts to assess the patient's status. She decides to consult with the dentist. Unlike Expert A, she does not explore possible correlates of the condition before referring.

Interviewer 2: OK you're making a face. What are you looking at?

Competent B: Six months previous perio charting. OK there is a large difference ... a big difference in the probing depths.

Interviewer 2: And how do you interpret that?

Competent B: Well, how do I interpret it, I mean I see a 5 here and now I see an 8.

Interviewer 2: What are you thinking?

Competent B: I'm thinking she's either got ... I'm thinking that medically something is going wrong here. It's not just periodontal problems. This is happening too quick to be just periodontally involved. OK I've done my probings. I've got x-rays. At this point I'm going to have the dentist take a look at this ...

Novice C fails to check for past periodontal charts or radiographs, merely continuing with a general exam of current conditions. She notes a gingival problem, but without comparing information over time cannot detect the rapid deterioration. Although attending to the patient's chief complaints, she has not placed them within a context of oral and medical health conditions.

Novice C: ... What did she it, esthetics, she, as far as does she have any restorations or crowns, things like that, that she is concerned about or does she have stain, or esthetically what is wrong with her teeth?

Interviewer 1: Old restorations ... Not attractive I believe, stained, dark.

Novice C: Okay, so I would think about replacing the restorations if we're talking about crowns. That would be something that she should discuss with the dentist.

Interviewer 2: Okay, and what would you do as you're doing this, would you ... what would you do, as you're doing this oral exam.

Novice C: I just would do a general exam and make note of any recession that she is sensitive to hot and cold in the upper right posterior. Look for faulty amalgams that maybe caused (inaudible) or some, in problem with her (inaudible) that she (inaudible) hot and cold or if she has recession or any decay, the general findings, the things that I would look for decay, leaky margins around restorations and decay under crowns. ... Just ask her these questions, when are ... or you're sensitive to hot and cold, making note of that so I can tell the dentist that, recording anything that I'm finding in her mouth that might be contributing to it.

Interviewer 2: Okay, do we have the results of the oral exam, I believe we do ...

Novice C: And it might indicate that she might need an x-ray on the upper right posterior.

Interviewer 2: You're looking in her mouth, so here is picture 2B in the front. Here is picture ... here's C and here's picture QA and that's what you see. Also the chart. She said that she would chart it, so let's give her the chart.

Interviewer 1: And this is the chart.

Novice C: I would assume that the sensitivity that she is having on the upper right, might be due to the faulty or leaking old broken restorations in number 3, indicating that she might need an extraction of that tooth. ... And the front of her ... as far as the anterior teeth, she ... I would probably recommend having restorations replaced on the lingual, that's at 9 or 10.

Interviewer 2: Okay.

Novice C: And then her best bet ... I'd hate to recommend anything because of the fact that she has got severe gingivitis, and so that problem would have to be addressed before we would think about any type of esthetics ...

Gathering and using information is especially important *in Patient assessment and evaluation,* which also appear as organizing categories.

Communication and language is related, since the use of domain terminology reflects integration of data features; that is, communication in terms of higher-level structures rather than specific findings is a cue that bits of information about the patient are being integrated into a domain-based knowledge structure.

Formulating and investigating hypotheses begins to take place even during initial phases of *Gathering and using information*. An expert hygienist has the knowledge base to construct a schema that integrates information about the client. These expert schemas suggest plausible underlying causes for the information received so far, and guide further information search and refinement of hypotheses about the client's condition. The expert reasons from observed data to possible underlying conditions, which in turn suggest further data that might be sought to sharpen the picture (Groen & Patel, 1988, call this pattern *forward reasoning* in medical diagnosis). Experts iteratively build and prune search trees. They solve problems more effectively than novices partly because they generate search trees that are more comprehensive and better matched to current information, partly because they are better at pruning the trees they have generated. Forward reasoning appears in this excerpt as an expert hygienist plans patient assessment after a brief introduction to the case:

Expert D: … OK, I'd look at the history first. I'd look for gastrointestinal illness that could be causing a bad taste, medications that cause dry mouth. OK. She has xerostomia, which is a dry mouth situation. It could be caused, sometimes a side effect of things like birth control pills, could be xerostomia, that could be a possibility of excessive thirst. Loss of weight. I'd question about the excessive thirst, see if that's related to medication.

Scripting behavior describes a phenomenon the Scoring Team saw at several phases of subjects' interactions with clients, including assessment, treatment planning, and patient education. A context, a situation, or a cue can activate a script. Novices learn to follow sequences of actions as "chunks," and carry them out with minimal tailoring to individual clients' needs. This is manifest when a novice produces a standard treatment plan or a patient education lecture with elements that are irrelevant or inappropriate for a particular client. Competent hygienists have integrated these chunks into more comprehensive knowledge structures, and vary them to some degree with the specifics of individual clients. Experts are able to construct assessment, treatment, and education plans according to individual client's needs, and are better able to monitor how they are proceeding and revise them as necessary.

The *Assessment* grouping comprises aspects of data gathering and interpretation that take place in patient assessment. The subcategory of *Global* assessment concerns the degree to which comprehensive schemas guide information gathering, to build a mental model of the patient. This finding is consistent with the research literature, which "support[s] a view of an expert as doing more inferential thinking and ending up with a more coherent model of the patient … In contrast, novice representations … [are] more superficial, fragmented, and piecemeal" (Lesgold et al., 1988, p. 317).

Excerpts from Case #5 illustrate this point with respect to oral examination. This patient shows signs of bruxism, or grinding teeth (recall Figure 4). Expert A recommends a TMJ (temporomandibular joint) evaluation along with the standard extraoral/intraoral examination, because occlusion problems can cause bruxism. Recalling his report of headaches, she establishes a possible connection among his lifestyle and medical history and her oral examination:

Expert A: As far as the headaches go, I would point out that he has so much attrition [worn surfaces], talk to him about the grinding. When do you seem to grind, do you notice yourself doing that?

Interviewer 1: Uh, no, my wife says it's at night, when I'm sleeping.

Expert A: She catches you doing it at night? Um, it wakes her up, is this every night?

Interviewer 1: Yeah, basically.

Expert A: You have some severe stress on that, and that could be part of the headache problem.

In contrast, Novice E performs the oral examination, but does not note the results. She moves on to radiography without commenting or acting on the results of the oral exam, which includes the photographs with clear (to an expert!) signs of attrition:

Interviewer 1: Okay, blood pressure is normal.

Novice E: Then I would do my internal and external oral exam.

Interviewer 1: All right, intraoral photograph of the mouth.

Interviewer 2: What are you thinking?

Novice E: Well, I'm thinking we need radiographs.

The *Interrelationships* and *Conditions* subcategories within *Patient assessment* focus on situations in which more specialized kinds of information must be gathered and integrated. Those in the *Procedures* subcategory concern how a hygienist gathers and uses information from standard sources, including background/medical information, dental charts, and oral, periodontal, and radiographic assessments. Novices show lapses in obtaining or interpreting information from any given source. Competent hygienists generally succeed at this level, but can fail to integrate information across the sources or to follow up on more subtle clues they hold.

The Scoring Team identified features of *Dental hygiene treatment planning* that reflect the completeness of the patient model the hygienist has constructed during patient assessment, as seen in the completeness and appropriateness of the resulting treatment plan or the referral to other professionals. Experts tended to perform more assessment in the initial part of the visit, a finding common to many domains, building appropriate mental models that lead them to more rapid and more appropriate action once they begin (Glaser & Chi, 1988). As a contrast, Novice E begins to follow a standard treatment plan of gross scaling, or removal of heavy deposits of calculus, despite contraindications in the oral photographs she has just seen:

Novice E: I would have that hand mirror out and I would have him look at it. I would point it out to him, explain to him what it is, and in lay terms, it's called tartar, in my terms it's called calculus, and explain to him what it is doing to his mouth.

Investigator 1: Is this what is causing all my bad breath?

Novice E: Um, yes.

Investigator 1: Can you get this off today?

Novice E: In 50 minutes, no. I would have to do quadrant scaling for gross calculus removal ...

Evaluation concerns monitoring the outcomes of dental hygiene treatments, and includes education as well as procedures. All the skills involved in *Assessment* and *Gathering and using information* are called upon, with an additional requirement of knowing usual outcomes of treatments under different patient conditions. Comparing data over time is especially important in evaluation.

Ethics is the remaining organizing category in Figure 5. Beyond substantive knowledge of dental hygiene, an important aspect of the profession is acting appropriately, legally, and morally. The Scoring Team noted that experts recognized ethical dilemmas as they arose and acted to resolve them. A key tenet of ethics in dental hygiene is recognizing the client's health status, orally and generally, as more important than the client's immediate desires if the two conflict. The hygienist must explain this to the patient as well as she can, and not carry out contraindicated procedures just to satisfy the client. Novices missed cues that suggested dilemmas to competent and expert hygienists. Here, the client in Case #7 wants Expert A to proceed with an invasive procedure, but she knows medical complications could arise if she does not refer the patient to her physician first.

Expert A: Okay, Mrs. White, the only problem I have even though your blood pressure is fine, um, everything you presented as a dental problem, is a good reason for you to be here, but before we do anything clinically, even though Dr. Goodfellow [her physician] never recommended it, I really think you should have that heart murmur checked out and let me get the periodontist in here so he can look at your health history as well. A murmur can be anything from a funny sound to a valve problem. It's not uncommon for a woman in her 30s to develop mitral valve prolapse, in which case you need an antibiotic before we can do anything extensive or invasive. So, I would have the periodontist come in, take a look at her medical history, speak to her, hopefully back me up and get her concerned enough to go to the HMO.

In this excerpt from Case #9, Competent hygienist F starts to say that she will do a general debridement because that is what the patient's condition indicates. She stops herself, though, saying that since his insurance won't pay for debridement, she'll do root planing instead; it is covered. She asks the client if he'd rather get an insurance refund or pay for debridement out of pocket. Later she struggles with the fact that he really needs a debridement but won't do it because of the insurance.

Interviewer 1: So what can you do today?

Competent F: Today, I'm going to do what's called a general debridement and get as much off of there as much as I can and then when we bring you back in a week's time, you're going to see a big difference in your gum tissues. But there's still a lot of stuff. I'm looking at your x-rays and I'll point out what the calculus look like. No, wait, let me rethink this cause you probably have insurance. If I do a debridement today, it won't be covered. You really should have root planing done [instead], unless you're willing to pay [for the debridement] out of pocket.

Interviewer 1: No.

[later in the case ...]

Competent F: This is fairly new to me right now … debridement. It just kills me to see all this stuff in here and not do a general debridement. It wasn't until recently that I found out that unfortunately the insurance won't cover root planing as well. This guy needs root planing. ... Too much stuff in there. It really would be best. Whether the insurance pays for it or not, it would be best to do a general debridement, get him back for four more visits after today.

This final excerpt is an example of an unrecognized ethical dilemma. Novice E agrees to treat the client within an hour, not realizing that his condition cannot be resolved in a single visit.

Interviewer 1: I'd sure like to know how quickly I can get out today? In a hurry.

Novice E: In a hurry?

Interviewer 1: Um-hum.

Novice E: Um, I'd ask for an hour, and I would take him ...

Interviewer 1: Can I get out of here in an hour?

Novice E: I'll give you an hour.

Interviewer 1: Okay, thanks, great that will work.

PHASE 5: DRAWING IMPLICATIONS FOR ASSESSMENT

The analyses described above were constructive and integrative, starting from observations of fairly unconstrained problem-solving and decision-making in settings. The sequence, the flow, and the direction of the subject's solution paths were determined by the subject herself. Such open-endedness is a key consideration in designing tasks for cognitive analysis. Since one cannot determine beforehand everything that will turn out to be important, one wants subjects to be free to take actions that might not have been anticipated.

The opportunity for discovery exacts a price. Collecting and analyzing these protocols from 31 hygienists consumed over a thousand hours of analysts' time. Even so, the results would serve poorly as assessment data about these 31 individuals. Explicit rubrics for eliciting and evaluating prespecified aspects of knowledge were essentially absent from the exercise (as appropriate to its purpose), so considerable examinee, case, and rater variance would appear in any scores derived from the protocols. The same procedure that is effective for generating an assessment framework would fail for assessing large numbers of examinees fairly and efficiently.

Tasks for large-scale, high-stakes assessment neither can be nor should be as open-ended. The values of validity and fairness demand that before one even presents a task, one can specify which inferences are to be made, what one considers evidence to support them, how observations will be interpreted as evidence, and how the task has been designed to evoke relevant evidence while minimizing the influence of irrelevant knowledge, skills, and experience. In contrast to cognitive task analysis cases or practice environments, assessment tasks must be carefully structured to evoke evidence about predetermined aspects of knowledge, individually or in combination, in terms of behaviors one knows how to interpret along predetermined lines. The tasks may still be very open-ended from the examinee's point of view; the tasks can be adaptive, interactive, amenable to many approaches, and offer a wide and realistic range of possible actions. The important constraints are the ones that the assessors impose upon themselves, namely the schema for evoking and evaluating examinees' behaviors. Only with this discipline can assessors design open-ended simulation tasks that they will know "how to score" ahead of time.

The cognitive task analysis provides copious information to help define such a framework for a simulation-based dental hygiene licensure assessment. It is clear by now that neither reporting, scoring, task design, or simulator development can be considered in isolation, so the discussions of implications for these models must overlap. We begin by focusing on the evidence model because the compendium of performance features was the chief product of the Scoring Team's analysis, and it provides the grist for defining the observable variables in the evidence model. We then consider implications for the student model, the task model, and the simulator in turn.

Implications for Evidence Models

The Structure of Evidence Models

The performance features from the cognitive task analysis characterize patterns of behavior that recur across cases and subjects, and capture differences among hygienists at different levels of proficiency. We propose to define a set of observable variables based on these features for use with all tasks in the assessment. The values of the observable variables encapsulate the evidence extracted from an examinee's individual and unique behavior (the transaction list of actions in the simulator environment), as it bears on the aspects of knowledge specified as student-model variables. Every task has a conformable evidence model, although one can create tasks that appear different to the examinee but use the same evidentiary structures (see Bejar & Yocum, 1991, on *task isomorphs*).

An evidence model also contains two kinds of relationships that are integral to evidentiary reasoning. First, *evidence rules* determine the values of observable variables from an examinee's work product. Table 2 gives an example (also see Clauser et al., 1997, and Gitomer, Steinberg, & Mislevy, 1995). Each of the observable variables $\underline{X}_1$-$\underline{X}_5$ can take a value of 0, 1, or 2 to characterize the quality of an aspect of performance in a given episode of performance in an assessment task. The column headed "Student-Model Parents" indicates the variables in a plausible student model that each might provide evidence about. The evaluation rules describe how features of the work product would be appraised to arrive at the values, in terms sufficiently abstract to be adapted to a range of particular cases. In a particular case, case authors would provide the details needed to carry out the evaluation in that case.

Second, a statistical model indicates how the observables in a given task depend on the student-model variables (the specification of which is addressed in a following section). These relationships are probabilistic, since novices sometimes make expert-like actions and vice versa; we seek to characterize accumulating tendencies. Familiar models from test theory can be employed at this stage, including item response theory, factor analysis, latent class models, and generalizations of them. In particular, we will be using Bayes nets to model the assessor's knowledge of the examinee's unobservable values of student-model variables, in light of evidence obtained thus far in terms of realized values of observable variables (Mislevy, 1994).

Considerations for Defining Observable Variables

What do the performance features of open-ended cognitive task analysis protocols suggest for defining observable variables for a simulation-based licensure exam? Proposing final configurations is beyond the scope of this article, but conclusions from the analysis can inform the decision. We conclude that we can define generally applicable observable variables from many of the performance features. Some performance features overlap enough to be combined. Others fall into groupings, as facets of the same kind of behavior. Still others will play roles in assessment design, but not as progenitors of observable variables.

Table 2

Examples of Evidence Rules

Observable variable	Range of values	Student-model parents	Evaluation rules (tailored to Task A)
X_1: Follow up on patient history	0: No follow-up 1: Follow up on documented exception conditions 2: Follow-up tailored to patient profile	Patient assessment Info gathering Procedures	**If** (patient history contains exception conditions **and** no follow-up questions are asked) **then** X_1 =0. **If** (patient history contains exception conditions **and** follow-up questions address only them) **then** X_1 =1. **If** (patient history contains exception conditions **and** patient profile contains associated features **and** follow-up questions probe connection) **then** X_1 =2.
X_2: Treatment consistent with patient stability	0: Issue of stability ignored 1: Treatment contraindicated by current patient stability 2: Treatment consistent with patient stability	Patient assessment Treatment planning	**If** (If blood pressure not taken) **then** X_2 =0. **If** (blood pressure taken and treatment A given **or** blood pressure taken and treatment A not given **or** blood pressure retaken and treatment A not given) **then** X_2 =1. **If** (If blood pressure retaken and treatment A given) **then** X_2 =2.
X_3: Compare/contrast of data from a given source over time	0: No comparison 1: Some comparison 2: Complete comparison	Patient assessment Info gathering Procedures	**If** (medical history 1 only) **then** X_3 =0. **If** (medical history 1 **and** medical history 2 **and** blood pressure medication **not** updated) **then** X_3 =1. **If** (medical history 1 **and** medical history 2 **and** blood pressure medication updated) **then** X_3 =2.

(table continues)

Table 2 (continued)

Observable variable	Range of values	Student-model parents	Evaluation rules (tailored to Task A)
X_4: Demonstrate understanding of effects of medication on oral/dental conditions	0: No demonstration 1: Superficial demonstration 2: Complete demonstration	Patient assessment Medical knowledge	**If** (no connection between dry mouth and medication) **then** X_4=0. **If** (connection made between dry mouth and blood pressure medication) **then** X_4 =1. **If** (connection made between dry mouth and anti-depressant) **then** X_4 =2.
X_5: Quality of treatment plan	0: Treatment inadequate/inappropriate 1: Treatment adequate/appropriate 2: Treatment optimal	Patient assessment Treatment planning	**If** (no fluoride treatment) **then** X_5 =0. **If** (fluoride treatment) **then** X_5 =1.

Figure 6 depicts performance features that could be refined into observable variables. Along the top are four broadly-defined performance features, which could constitute a small collection of generally applicable observable variables: *Assessment*, *Problem formulation*, *Treatment planning*, and *Ethics*. Below them are finer grained categories. For example, the two subcategories under *Assessment* concerning specific sources of information and circumstances for extracting information. Observable variables could be formally defined from these more narrowly-defined performance features to constitute a larger collection of generally applicable observable variables. Table 3 summarizes performance features at this level that could serve as the basis of observable variables, and gives examples of how they might be used.

We should note in passing that potential observable variables such as *Patient assessment* and *Treatment planning* have names that could also be used to define student-model variables. These two kinds of variables are distinguished by their roles. The observable variables would characterize features of specific task performances, while student-model variables would characterize features of a candidate's skill and knowledge. An observable variable called *Patient assessment* associated with a particular task would be used to describe the character or the quality of the patient assessment procedures employed in a particular response to that task. A student-model variable called *Patient assessment* would be used to characterize a candidate's propensity to carry out comprehensive patient assessments across the range of cases that comprise the domain.

An assessment system with observable variables only at the coarse grain-size would rely heavily on task-specific rubrics to map performances into its sparse framework. The distinctions that appear in the more narrowly defined performance features would be used to inform the evidence rules for parsing transaction lists and evaluating observable variables for specific tasks. The alternative is to work with a greater number of observable variables, corresponding to the more narrowly-defined performance features. Evidence models will be more complex, but writing evidence rules for individual tasks would be easier. Furthermore, capturing and distinguishing features of performance at this level of specificity maintains the possibility of including student-model variables at this same grain-size for inferential reporting or feedback to candidates.

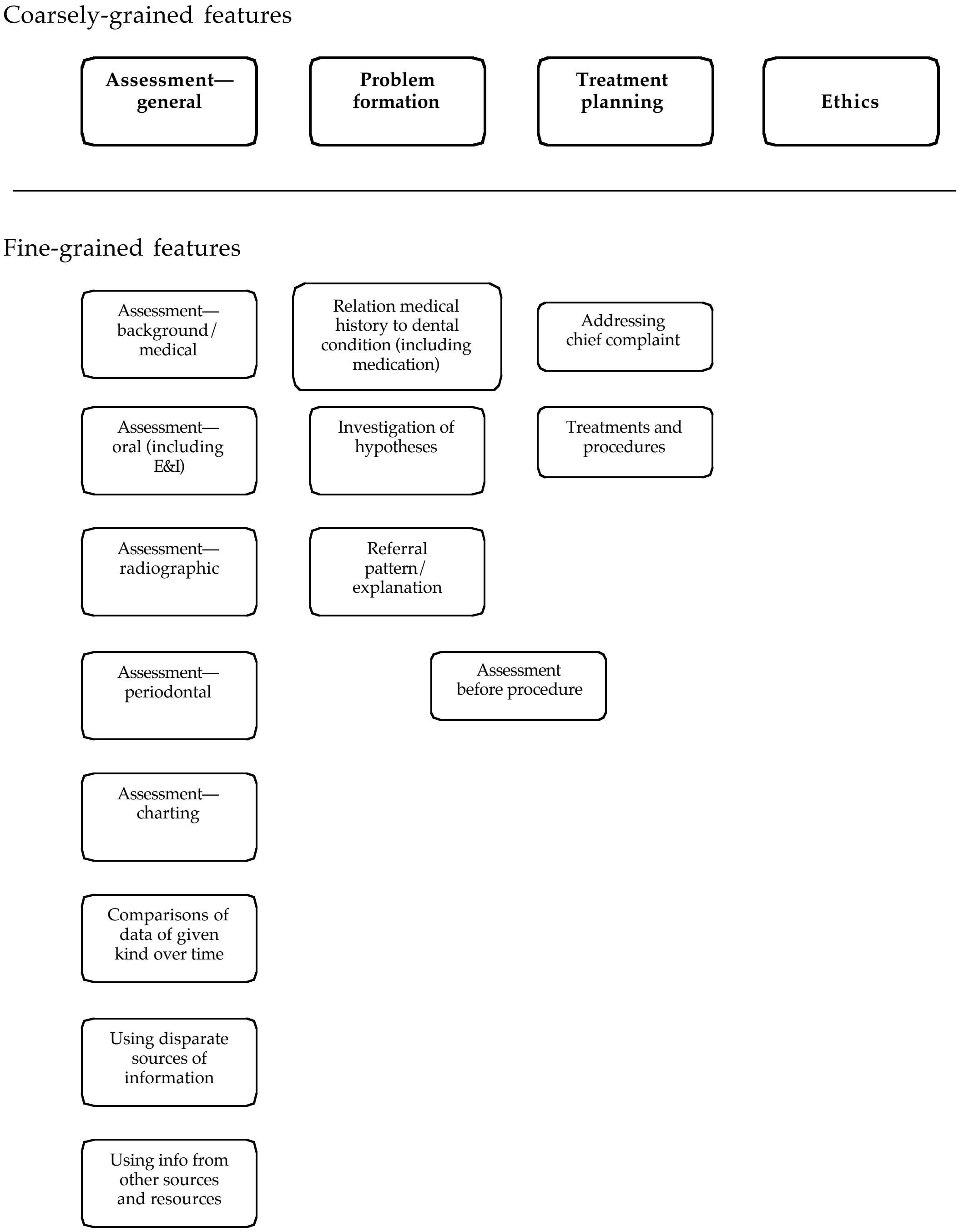

Coarsely-grained features
Assessment—general
Problem formation
Treatment planning
Ethics
Fine-grained features
Assessment—background/medical
Relation medical history to dental condition (including medication)
Addressing chief complaint
Assessment—oral (including E&I)
Investigation of hypotheses
Treatments and procedures
Assessment—radiographic
Referral pattern/explanation
Assessment before procedure
Assessment—periodontal
Assessment—charting
Comparisons of data of given kind over time
Using disparate sources of information
Using info from other sources and resources

Figure 6. Performance features that are candidates for observable variables.

Table 3

CTA Performance Features and Implications for Defining Observable Variables

Broad categories of performance features	Finer-grained categories	Potential observable variables	Examples of situations to evoke evidence
Patient assessment	Background/ medical	Quality of follow-up on patient history; medical history; dental history; psycho-social factors.	Patient situation indicates asking for background information, medical history, or changes in medical condition.
	Oral (extraoral and intraoral)	Thoroughness of exam(s); correctness of recorded of findings.	Patient situation indicates conducting extraoral and/or intraoral exam; requirement to record findings.
	Radiographic	Use of radiographs to confirm/deny physical findings; interpretation of features of radiographs.	Patient condition indicates a need to request existing radiographs, or a need to order appropriate new radiographs.
	Periodontal	Use of periodontal findings to confirm/deny physical findings; interpretation of radiographs.	Patient condition indicates a need to request previous periodontal exam findings, or to conduct new exam
	Charting	Accessing/creating/interpreting charting information.	Patient situation indicates accessing existing charts; creating or augmenting charts with new exam results; examinee is asked to summarize information from a complex chart.
	Comparisons of data over time	Compare perio/radiographic/oral data to determine progress or detect anomaly.	Need to access past charts to compare with new findings; opportunity to record new findings for future comparisons.
	Using disparate sources of information (i.e., data integration)	Combine information from periodontal and radiographic data; from current exam findings and databases.	Clues to dental condition are distributed across periodontal, oral, and radiographic data, which must be integrated to detect patient status.

(table continues)

Table 3 (continued)

Broad categories of performance features	Finer-grained categories	Potential observable variables	Examples of situations to evoke evidence
Patient assessment (continued)	Using information from other sources and resources	Appropriateness of seeking information from dental consultant, medical reference.	Need to request treatment information from a periodontist regarding periodontal care; patient presents with new medication, about which dental implications must be sought from a drug database
Problem formation	Relating medical history to dental conditions	Degree to which observations bearing on medical conditions prompt focus on oral/dental conditions, and vice versa.	Patient medical history has features that contraindicate standard examination procedures; findings from oral exam suggest unrecognized change in medical status.
	Investigation of hypotheses	Demonstration of pursuit of additional key information, given information so far; presence and appropriateness of modifications of hypotheses in response to new information.	Requirement to generate hypotheses after specific pieces of information are provided; requirement to provide rationale for further information requests.
	Referral pattern/explanation	Appropriateness of and rationale for referrals.	Opportunity available to make referral to a physician, medical specialist, dental specialist, pharmacist, social worker, etc.—sometimes necessary, other times not; requirement to provide rationale to patient.
Treatment planning	Assessment before procedure	Degree and appropriateness of assessing patient medical and dental status before planning treatment.	In new appointment, opportunity presented to make (appropriate) inquiries about changes in medical/dental history before beginning previously arranged procedure.
	Addressing chief complaint	Degree to which patient's chief complaint is addressed in assessment/treatment plan, as appropriate.	Chief complaint presented which requires follow-up of medical history; chief complaint presented which must be postponed until previously unrecognized dental problem is addressed.

(table continues)

Table 3 (continues)

Broad categories of performance features	Finer-grained categories	Potential observable variables	Examples of situations to evoke evidence
Treatment planning (continued)	Treatments and procedures	Appropriateness of procedures in plan; appropriateness of sequence; carrying out of procedures in accordance with plan.	Opportunity present to performs procedure (scaling, root planing, etc.); opportunity to schedule follow-up in appropriate or inappropriate period of time; opportunity to check progress at follow-up; opportunity to perform the next procedure or revise plan in light of new information.
Ethics		Recognition of ethical dilemma; addressing recognized dilemma in a professional manner.	Patient provides information belied by exam findings; evidence of inappropriate care by previous hygienist; conflict between patient oral health needs and insurance regulations.

As mentioned above, *Patient assessment* encompasses a set of more narrowly-defined performance features that concern circumstances for extracting information: comparing data from a given source over time, using disparate sources of information, and using other information sources and resources. Tasks would be designed to evoke evidence about these features by controlling the information that resides in various data sources, how it interrelates, and what one can find in a transaction list that provides evidence about an examinee's proficiencies. With such tasks, the features that must be extracted from a transaction list will concern whether the examinees' actions call up the data sources and whether subsequent actions reflect use of the information they contain.

In the cognitive task analysis, the interviewers frequently asked the subjects "What are you thinking now?" or "What does that tell you?" Their answers provided direct evidence about the subjects' hypotheses and their reasons for carrying out actions and requests for further information: the performance features *Problem formation* and *Investigation of hypotheses*. Direct evidence of what subjects are thinking is not as easy to obtain in large-scale assessment. In a large-scale simulation-based assessment, one might confound evidence about patient assessment and hypothesis formulation by allowing an examinee to seek information with minimal structuring or guidance, so evidence about hypotheses arrives only implicitly through features of the search path. This is realistic, but it provides relatively little information about the examinee's inferences and hypotheses. Alternatives are to require the examinee to indicate provisional hypotheses at prespecified points in the task, and to solve short cases that require only hypothesis generation from specified information. An operational assessment can include cases that mix evidentiary ingredients in these different ways to cover a spectrum of realism and targeted evidentiary focus.

Referral pattern and *Explanation for dental referral* could be combined into a single observable variable. How can we structure tasks to provide evidence of this kind? Toward the less constrained end of the "openness" spectrum, an affordance for referrals (medical, dental, pharmacological) could always be available to a candidate, to acquire data or to end a case. A candidate's use of this capability would constitute the evidence. Toward the constrained end of the spectrum, a case could explicitly ask at a specified point whether the candidate would recommend a referral. Either way, we could obtain direct evidence by requiring a rationale for the referred professional. The nature of the referral would depend on the affordances the simulator could support; in order of increasing realism and expense but less predictable evidentiary value: marking a checklist, choosing from a long list of possibilities, or providing a natural language explanation.

Treatment planning appears as a more broadly-defined performance feature that encompasses more focused performance features: addressing the chief complaint, treatments and procedures, and degree of assessment and planning before carrying out a procedure. The last, while describing a characteristic of treatment planning, also provides evidence about skill in patient assessment. In a system that included both *Patient assessment* and *Treatment planning* as student-model variables, such an observable variable provide evidence about both.

Ethics contains no more narrowly-defined performance features, so if DISC included an *Ethics* variable in its student model, the present analysis provides no distinctions for finely-grained and coarsely-grained approaches to defining observable variables. Values for an *Ethics* observable variable would be derived with evidence rules that scanned a transaction list for actions that suggest an ethical dilemma designed into the case is (a) detected and (b) dealt with at one of a number of designated levels of quality.

Other performance features would not evolve into observable variables themselves, but would play other roles in assessment design. *Patient assessment with a difficult performance situation feature* signals a task feature that will make assessment more difficult. *Assessment with a complex medical condition* requires the usual expert-like behavior for proficient performance, but additionally places heavier demands on medical knowledge. A task with a complex medical condition would provide evidence about student-model variables concerning patient assessment (background/medical), hypothesis formulation, and treatment planning.

Use of scripts was explicit in the talk-aloud protocols, but it would be less clear if observations consist of only a list of actions. To incorporate this feature into a structured assessment, one would construct tasks that would call for a straightforward procedure if not for a complicating factor, contra-indication, or anomalous potential finding. “Scripting” is then probably occurring only if information is not sought, or is sought and found but subsequently disregarded. Such patterns would be captured in evidence rules that evaluated observable variables such as *Assessment before treatment* and *Seeking and using information*.

Vocabulary and language use related to knowledge retrieval concerns the words and structures a hygienist used to describe findings and procedures in assessment, and marked developing expertise in the open-ended protocols. The difficulty of interpreting natural language responses may preclude productive use of language in a large-scale high stakes assessment. More directed (and less natural) alternatives include having examinees choose among different ways of expressing an hypothesis and filling in a partially formed patient education script.

Considerations for Defining Evidence Rules

We have discussed evidence rules in general terms up to this point. Three levels of evidence rules can actually be distinguished. At the lowest level are those that simply parse the raw work product to extract relevant elements of the production and strip away irrelevant elements. At the next level are those that identify the presence, absence, or extent of predesignated meaningful features. At the third level are rules that combine those features to yield values of specified observable variables. It is examples at this third level that appeared in Table 2. Most interesting measurement issues arise at this level, and they can profitably be discussed even before many specifics of implementation have been determined.

The most common form of “scoring” in performance assessment has been to ask human experts to rate the overall quality of the examinees’ performance. The costs, operational difficulties, and uncertainties associated with human rating motivate interest in automated evaluation of performances (Braun et al., 1990). High correlations between automated summary evaluations and human ratings are taken as evidence of success: the implied objective being to duplicate, at lower costs and with greater reliability, what the experts are doing. But simply replacing humans with computer programs in an existing assessment system forgoes the opportunity to design a system that capitalizes on the different strengths of the available components and effectively integrates them.

The research on behavioral decision-making, extending back to Meehl's (1954) classic *Clinical versus statistical prediction*, argues that the best way to combine the strengths of humans and computers is to lean more heavily on human experts to *identify* salient features and on empirical models to *combine* their import. This argues for using multiple observable variables to summarize distinguishable features of performance, even if only a single overall measure will ultimately be required. The job of combining disparate forms of evidence, so difficult for humans, can be dealt with in a formal and principled statistical framework. Rules (mechanical or human) to extract and evaluate features should thus focus on more narrowly-defined aspects of performance, and use relatively simple rules of combination to map them into common observable variables. This works best if tasks are designed expressly to place the examinee in settings where actions can be evaluated along predetermined lines, a job for which experts' insights are of signal value.

Implications for the Student Model

Student-model variables correspond to what Messick (op cit.) called "the complex of knowledge, skills, or other attributes [that] should be assessed." The student model itself is used to accumulate evidence about them, and characterize the assessor's current knowledge about their inherently unobservable values. The purpose of an assessment drives the nature, number, and grainsize (i.e., level of detail) of variables in the student model. Student-model variables persist over the course of tasks in an assessment, evidence from each task in turn used to update them. The basic principle for defining student-model variables for an assessment is that they must be consistent with the purpose of the assessment. This overarching goal has the following corollaries.

Student model variables should be consistent with substantive domain. In several ways, the initial stages of the cognitive task analysis ensured that performance features extracted from the protocols would be substantively relevant and appropriately targeted. The Scoring Team started by specifying the job domain: the behaviors, knowledge, contexts, and activities that characterize the profession, including markers of developing proficiency. From this foundation they created a set of tasks that would (a) be representative of the domain, (b) exemplify valued work, (c) focus on decision making and problem solving, and (d) be likely to elicit different behavior from hygienists at different levels of proficiency. This process ensures that observations will hold evidentiary value for inferences about examinees' skill and knowledge cast within this general framework. It does not, however, determine the exact nature or number of student-model variables.

Student model variables should be consistent with the required inferences. The specifications map the scoring team created, the organization of dental hygiene instructional material (e.g., Darby & Walsh, 1995), and existing dental hygiene licensure assessments overlap considerably as to the major components of interactions with patients; they are patient assessment, data interpretation, planning hygiene care, implementation, evaluation, and communication. Most of these categories are also aligned with the performance features extracted from the cognitive task analysis protocols. With the exception of *implementation of procedures*, which is not a focus of the proposed assessment, they are strong candidates for student model variables. Figure 7 is a plausible configuration for a student model for the proposed assessment (a fragment of a Bayes net, including student model variables but not observable variables). The student-model variables in such a model would be used to accumulate evidence across tasks, from the finer-grained observable variables capturing evidence from specific tasks (a conformable fragment of a Bayes net including observable variables for the task at hand would be "docked" with the student model for this purpose; see Almond & Mislevy, in press). These variables would then provide for an overall summary measure of performance and for cognitively-relevant feedback on performance across cases.

These construct-based summary measures and feedback are *inferential* uses of data. They are mediated through student-model variables, and as such merit scrutiny of reliability. An example is, "You tend to do poorly in situations that require comparing patient data across time." In contrast, *descriptive* feedback simply reports what an examinee did in particular tasks. An example is, "In Case #2 you missed the rapid bone loss in the past six months." The amount and detail of *descriptive* feedback in a simulation-based assessment can be as rich as is useful to candidates, as long as it does not imply how an examinee would do on other tasks, offer suggestions for study, or indicate suitability for licensure.

Student model variables should be measured with sufficient accuracy for the decisions or feedback they support. While it is clear that student-model variables defined at the grainsize illustrated above could be used for inferential feedback along cognitively-relevant lines, how would they provide for an overall pass/fail decision—a decision that is, after all, the primary purpose of the proposed assessment? A natural approach is to provide for an arbitrary function of the final distribution of the student-model variables after assessment. The combining function depends on policy as much as statistics. Should the student-model variables be weighted equally, or differentially in proportion to their adjudged importance? Should the variables be combined with a compensatory function such as a linear combination, so that good performance on some aspects of problem-solving makes up for bad performance on others, or with a multiple-hurdle rule so that a candidate must perform at some specified minimal level on all aspects of knowledge to pass?

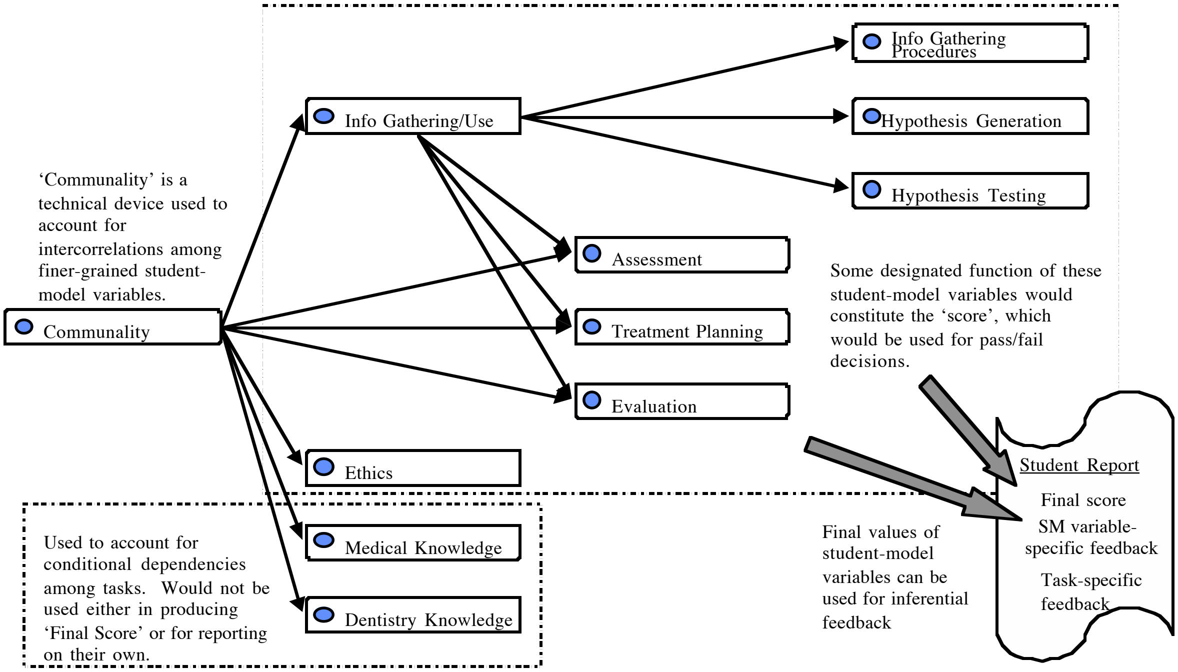

Info Gathering Procedures
Info Gathering/Use
Hypothesis Generation
Hypothesis Testing
'Communality' is a technical device used to account for intercorrelations among finer-grained student-model variables.
Assessment
Treatment Planning
Evaluation
Communality
Some designated function of these student-model variables would constitute the 'score', which would be used for pass/fail decisions.
Ethics
Student Report
Final score
SM variable-specific feedback
Task-specific feedback
Used to account for conditional dependencies among tasks. Would not be used either in producing 'Final Score' or for reporting on their own.
Medical Knowledge
Dentistry Knowledge
Final values of student-model variables can be used for inferential feedback

Figure 7. Annotated student model.

A standard-setting study helps to answer these questions. Subjects in the neighborhood of the decision are administered the new assessment, and experts, using either external criteria or a holistic evaluation of assessment performance, designate each subject as a *pass* or a *fail* (Livingston & Zieky, 1982). Functions that model the overall expert evaluation in terms of student-model variable values can then be explored. Reliability issues abound in performance assessments such as the proposed simulation-based licensure exam. We will study measurement error with respect to both reliability of pass-fail decisions and alternate forms of the assessment.

A common finding in performance assessment is that the number of cases required for accurate measurement can be surprisingly high (Linn & Burton, 1994), the so-called low generalizability problem. The major issue is case specificity: Which cases are hard and which are easy varies considerably from one examinee to the next, so quality of performance on one task says little about performance on the next. This is due partly to the breadth of domains that have been studied and partly to the complexity of the performances that have been mapped to summary scores. The approach proposed here aims to ameliorate this effect in two ways:

- **Integrated assessment design**. The performance tasks Linn and Burton (op cit.) describe and others in the studies they cite were assembled under a task-centered approach to assessment design: Tasks are created because they are interesting and apparently relevant, with the presumption that they can somehow “be scored” and that an overall propensity to “score high” will be a coherent construct. But the more complex an assessment is—the more multivariate the student-model, the more interactive and multifaceted the performance—the less likely a task-centered approach is to succeed. By designing tasks expressly to focus their evidentiary value on specified targets of inferences, in ways we know beforehand how to interpret, we increase construct relevant variance and reduce construct irrelevant variance.

- **Disentangling sources of variation**. Performance assessments have low generalizability when different tasks in a heterogeneous domain tap different mixes of skill and knowledge, producing “noise” when only overall proficiency is captured (Wiley & Haertel, 1996). Systematic performance differences across tasks can be captured as construct-relevant “true” variance with a finer grained student model if the relationships between task demands and requisite finer-grained aspects of knowledge are modeled appropriately, as we intend. Suppose that a

> student-model includes variables for both Patient Assessment and Medical Knowledge, and the latter is a parent of just those observables that require medical knowledge. Weak patient assessment in cases requiring more medical knowledge and better assessment in cases requiring less will be captured as a low value for Medical Knowledge and a higher value for Patient Assessment. This uneven profile does, of course, introduce variance into a summary score. However, principled apportionment of possible explanations in terms of the profile of finer-grained student-model variables both reduces the uncertainty and provides an opportunity for feedback as to where and how, in substantively meaningful terms, the uncertainty arises.

Student model variables should be consistent with evidence models, task models, and assessment assembly specifications. Just as it is ineffective to design tasks in isolation and hope someone will figure out how to score them, it is ineffective to list student-model variables (or equivalently, feedback and reporting categories) without considering how to gather and interpret evidence that informs them. If we want to make an inference about some aspect of examinee skill or knowledge, we must (a) build *evidence models* that indicate what we need to observe in order to update our knowledge about that aspect of skill or knowledge; (b) design *task models* that indicate just how tasks are described, constructed, and aimed to ensure that opportunity to obtain observations that bear on that variable; and (c) write *assessment assembly specifications* that indicate how tasks with different evidentiary value for different variables are to be combined to produce an optimal mix of evidence, including evidence about the targeted variable.

Implications for Task Models

Task model variables concern the characteristics of tasks that were discussed above less formally as task features. The DISC cognitive task analysis informs the definition of tasks and task models in two ways. First, the set of analysis cases themselves may legitimately be viewed as seeds for potential task models, because they were designed in full consideration of an explicit articulation of a body of relevant evidence, namely, the expert committee's specification of the roles, materials, and conditions of dental hygiene practice. Second, whether or not the cognitive task analysis tasks are used further, the findings should ground the design of task models and simulation capabilities.

Before discussing implications for task models and the simulator, we should point out how closely the two are interrelated. Task models are templates for designing individual tasks. The simulator is the supporting environment in which examinees interact with these tasks. Therefore, basic simulator requirements should be derived from resource and processing requirements across all task models. Task and simulator design are not determined by evidentiary requirements alone, however. They also depend on the affordances and constraints imposed by the operational environment (platform, time available, pricing, etc.). The advantage of evidence-centered task and simulator design is that their features can be evaluated in terms of the evidence they provide.

The structure of task models. A task model is a collection of task model variables used to characterize and create individual tasks. Its definition includes its scope, or the range of values each variable may take in tasks written under that model. If the range of values for a Periodontal Status task variable is 1-to-5, a task model for generating cases that afford the opportunity to demonstrate periodontal expertise might constrain its values to 3-to-5. Task models facilitate writing tasks, because values of task model variables can be inherited from a common structure, which has been configured to achieve certain objectives. A task model's scope may be based on evidentiary focus (e.g., tasks about dental hygiene treatment planning), substantive characteristics (e.g., tasks involving only pediatric patients), or response features (e.g., tasks requiring only multiple-choice responses).

The features articulated in the cognitive task analysis tasks inform the definition of variables for licensure assessment task models. Table 1 offers suggestions for task model variables that are derived from the cognitive task analysis cases. The table entries are only suggestive of the full ranges of variables and of values for those variables. In addition to task model variables motivated by the analysis, the following kinds of task model variables will also be required for a simulation-based licensure exam:

Level of difficulty/Value of information. This specifies a level of difficulty for a given task, or the value of information a task provides for discriminating at a given level of proficiency. Collis et al. (1995) and Dennis, Collis, and Dann (1995) use the term *radical* to describe task model variables having values that affect task difficulty (e.g., medical condition), and *incidental* for task model variables that do not (e.g., patient name). It may be desirable to specify multiple levels of difficulty, relative to different student-model variables, to guide task selection in multivariate adaptive testing in the future.

Focus of evidence. These task model variables are used to specify what the task is about; that is, its evidentiary focus. Values for this variable can be expressed in terms of student-model variables and/or observable variables, at their grain-size or finer. Sample values for three hypothetical tasks indicating the sets of student model or observable variables about which each task provides evidence:

- assessment, demonstrates understanding of effects of medication on oral/dental conditions;
- treatment planning, demonstrates formulation of appropriate patient education, pediatric;
- treatment planning, evaluation.

Pacing. Pacing concerns how much time the examinee will have to complete the task. It is important to define how long the exam will run so that tasks (cases) can be designed with time constraints in mind. Once the student model variables have been determined, a decision can be made about how many of which kinds of cases are required for reliable evidence for any particular student model variable or the summary score.

Implications for the Simulator

The basic requirements of the simulator—that is, what is necessary for a user to accomplish given work in a specified domain, independent of any particular solution or implementation—should be derived from the resource and processing requirements of all the task models that will generate tasks to be used in the simulator environment. The term *resources* will encompass all the objects in the simulator's user interface that an examinee can interact with, and the means for doing so; that is, all those materials, agents, actions, and procedures put at the examinee's disposal to perform a simulator-based task. *Processing* will encompass the means for producing the actual results of interaction; that is, what is generated when the examinee uses resources.

Resources

Resources constitute the materials, actions, and procedures that provide an appropriate general context for task performance (e.g., the standard appurtenances of a dental hygiene operatory), and specific support for individual tasks (e.g., looking up the side effects of a medication in a specific reference book). Resources incorporated into the cognitive task analysis tasks generally fell into the categories of documents (text and graphics), people, equipment, searching, communicating, using, and interpreting. These tasks used a noteworthy variety of documents, and emphasized the search for and the interpretation of information from them. This finding argues for task authoring tools that can easily incorporate new types of documentation and search procedures into the simulation environment (e.g., searching literature and databases). Further, the simulator should support task execution in the following ways:

- capability to state hypotheses (as discussed above, by means that can range from multiple-choice, to choosing from an extended list, to short natural-language responses);

- capability to acquire resources interactively (knowing how to interpret given information is not enough; a hygienist must know what to look for, where to look, and how to find it);

- capability to make explicit causal or explanatory connections between various information and observations: e.g., among medical, social, occupational, and behavioral characteristics; symptoms and complaints; oral and dental conditions; and elements of treatment plans.

Processing

The experts who administered the cognitive task analysis interviews had to produce results of particular actions or decisions for the subjects. Two analogous requirements for the simulator follow: capabilities to (a) configure available resources dynamically in response to actions or conditions, and (b) produce changes in patient dental, oral, and periodontal status dynamically as a function of specific actions (or lack thereof), medical conditions, elapse of time, or some combination.

During the analysis we observed that the evidentiary value of responses was diminished when the interviewers (substitute “the simulator”) “kept subjects on path” with certain kinds of feedback. Examples included presenting information without a request; providing interpretations of information or observations; and indicating that a particular resource is unavailable or not allowed.

Most importantly, simulator requirements for the licensure examination depend on the final student model. Target inferences determine evidence requirements, which determine task features, which determine simulator support. The advantage of an evidence-centered approach to simulator design is that functionality can be rationalized in terms of the value of the evidence it produces. Because of the anticipated complexity of the simulation environment, it would be wise, as part of the simulator design process, to verify that users engaged in simulation-based cases can accomplish their cognitive goals with available simulator functionality and without undue difficulty. Available cognitively-oriented analytic procedures can be used for this purpose (see Steinberg & Gitomer, 1993, for use of GOMS methodology in designing the Hydrive interface).

Next Steps

The cognitive task analyses lay the groundwork for defining the assessment variables and models for a coherent and principled simulation-based assessment of problem-solving and decision-making in dental hygiene. DISC will make a final determination of the student-model variables, in line with the intended use of the assessment and the evidence requirements the analysis revealed. This decision sets the stage for formally defining observable variables, as to their grain-size and the data to be generated by candidates' interactions with the simulator.

Final definitions of task models will require determinations of assessment constraints, student model, and evidence models. The simulation capabilities developed by DISC to date give rise to task model variables that control the presentation, structuring, and content of the cases. The cognitive task analysis has uncovered requirements for additional task model variables that concern evidentiary properties of tasks. Still other task model variables will be needed to assemble tasks into assessments. As DISC design teams begin to develop task models that (a) support inferences about the student model variables they decide to target and (b) yield values of observable variables that contain evidence about those student-model variables, they will examine the simulator affordances, constraints, and interfaces needed to produce that evidence. They can then tune and extend the capabilities of the current simulator in line with the evidentiary needs they thus define.

A FINAL COMMENT

Problem-solving in the real world is rarely circumscribed as neatly as in assessment tasks. Problems are unique, woven into social contexts, perceived in terms personal to the problem-solver. A problem might be described in terms of myriad features, as could the problem-solver's evolving encounter with the situation. Somehow, informally, intuitively, we make inferences about the skills and knowledge of the people with whom we interact.

This scene contrasts most obviously with standardized testing by the realism of the situations and the richness of the data. Tasks in standardized tests are encapsulated and observations are spare mainly because historically, we could not handle more. Computers and simulation capabilities shatter this barrier—only to reveal a new one: just how to make sense of "rich and realistic data." We now recognize a more subtle difference between everyday inference and standardized tests. Standardized tests have proven useful despite the sparseness of their data because of the methodologies that have evolved to put that data to use—to guide its collection, to summarize its value, to characterize its accuracy, to critique its effectiveness; to gather, to encapsulate, to communicate its evidentiary import over time, across distance, between people.

While specific methodologies of familiar tests may fall short for transforming computer-based simulation capabilities into valid assessments, the same evidentiary principles upon which they are based (Schum, 1994) can ground methods that are up to the task. The challenge is to create structures for designing simulation-based tests, making principled observations, and drawing defensible inferences. The response, illustrated in this article in the context of problem-solving in dental hygiene, is a coherent evidentiary framework for modeling targets of inferences, aspects of observations that provide evidence about them, and features of situations that evoke that evidence.

REFERENCES

Almond, R. G., & Mislevy, R. J. (in press). Graphical models and computerized adaptive testing. *Applied Psychological Measurement.*

Almond, R. G., Mislevy, R. J., & Steinberg, L. S. (1997). *Task design, student modeling, and evidentiary reasoning in complex educational assessments.* Poster presentation for the Section on Bayesian Statistical Science at the Annual Meeting of the American Statistical Association, Anaheim, CA, August 10-14, 1997.

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.

Bejar, I. I., & Yocum, P. (1991). A generative approach to the modeling of isomorphic hidden-figure items. *Applied Psychological Measurement, 15*(2) 129-137.

Braun, H. I., Bennett, R. E., Frye, D., & Soloway, E. (1990). Scoring constructed responses using expert systems. *Journal of Educational Measurement, 27,* 93-108.

Clauser, B. E., Subhiyah, R., Nungester, R. J., Ripkey, D., Clyman, S. G., & McKinley, D. (1995). Scoring a performance-based assessment by modeling the judgments of experts. *Journal of Educational Measurement, 32,* 397-415.

Collis, J. M., Tapsfield, P. G. C., Irvine, S. H., Dann, P. L., & Wright, D. (1995). The British Army Recruit Battery goes operational: From theory to practice in computer-based testing using item-generation techniques. *International Journal of Selection and Assessment, 3*(2).

Darby, M. L., & Walsh, M. M. (1995). *Dental hygiene: Theory and practice.* Philadelphia, PA: W. B. Saunders.

Dennis, I., Collis, J., & Dann, P. (1995). *Extending the scope of item generation to tests of educational attainment.* Proceedings of the International Military Testing Association, Toronto, October, 1995.

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.

Gitomer, D. H., Steinberg, L. S., & Mislevy, R. J. (1995). Diagnostic assessment of trouble-shooting skill in an intelligent tutoring system. In P. Nichols,

S. Chipman, & R. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 73-101). Hillsdale, NJ: Lawrence Erlbaum Associates.

Glaser, R., & Chi, M. T. H. (1988). Overview. In M. T. H. Chi, R. Glaser, & M. J. Farr (Eds.), *The nature of expertise* (pp. xv-xxxvi). Hillsdale, NJ: Lawrence Erlbaum Associates.

Groen, G. J., & Patel, V. L. (1988). The relationship between comprehension and reasoning in medical expertise. In M. T. H. Chi, R. Glaser, & M. J. Farr (Eds.), *The nature of expertise* (pp. 287-310). Hillsdale, NJ: Lawrence Erlbaum Associates.

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

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: Lawrence Erlbaum Associates.

Linn, R. L., & Burton, E. (1994). Performance-based assessment: Implications of task specificity. *Educational measurement: Issues and practice, 13* (1), 5-8.

Livingston, S. A., & Zieky, M. A., (1982). *Passing scores: A manual for setting standards of performance on educational and occupational tests*. Princeton, NJ: Educational Testing Service.

Meehl, P. E. (1954). *Clinical versus statistical predication: A theoretical analysis and a review of the evidence*. Minneapolis: University of Minnesota Press.

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. (1992). 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., & Gitomer, D. H. (1996). The role of probability-based inference in an intelligent tutoring system. *User-Modeling and User-Adapted Interaction, 5,* 253-282.

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

O'Neil, H. F., Allred, K., & Dennis, R. A. (1997). Validation of a computer simulation for assessment of interpersonal skills. In H. F. O'Neil (Ed.), *Workforce readiness: Competencies and assessment* (pp. 229-254). Mahwah, NJ: : Lawrence Erlbaum Associates.

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

Schneider, B., & Konz, A. M. (1989). Strategic job analysis. *Human Resources Management, 28,* 51-63.

Steinberg, L. S., & Gitomer, D. H., (1993). Cognitive task analysis and interface design in a technical troubleshooting domain. *Knowledge-Based Systems, 6,* 249-257.

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