Leverage Points for Improving Educational Assessment

Deliverable – September 2000

Project 3.2 Validity of Interpretations and Reporting Results—
Evidence and Inference in Assessment

Robert J. Mislevy, Project Director
Educational Testing Service, Princeton, New Jersey

U.S. Department of Education
Office of Educational Research and Improvement
Award #R305B60002

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Center for the Study of Evaluation (CSE)
Graduate School of education & Information Studies
University of California, Los Angeles
301 GSE&IS, Box 951522
Los Angeles, CA 90090-1522
(310) 206-1532

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

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

TABLE OF CONTENTS

ABSTRACT **1**

INTRODUCTION **2**

EVIDENCE-CENTERED ASSESSMENT DESIGN **3**
Reasoning from Complex Data *3*
Evidentiary Reasoning and Assessment Design *4*
The Student Model *5*
Evidence Models *11*
Task Models *18*

LEVERAGE POINTS FOR IMPROVING ASSESSMENT **22**
Leverage Points for Psychology *23*
Leverage Points for Statistics *28*
Leverage Points for Technology *31*

CONCLUSION **37**

REFERENCES **38**

LEVERAGE POINTS FOR IMPROVING EDUCATIONAL ASSESSMENT[1]

Robert J. Mislevy, Linda S. Steinberg, & Russell G. Almond

Educational Testing Service, Princeton, NJ

Geneva D. Haertel & William R. Penuel

SRI International

Abstract

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

[1] This paper was prepared for Technology Design Workshop sponsored by the U.S. Department of Education, held at Stanford Research Institute, Menlo Park, CA, February 25-26, 2000. We Thank Barbara Means for helpful comments on an early draft.

INTRODUCTION

Interest in complex and innovative assessment is expanding nowadays for several reasons. With advances in cognitive and educational psychology, we are coming to better understand how people learn, how they organize knowledge, how they put it to use (Greeno et al., 1997). This broadens the range of things we want to know about students, and possibilities for we might see to give us evidence (Glaser, Lesgold, & Lajoie, 1987). We have opportunities to put new technologies to use in assessment, to create new kinds of tasks, to bring them to life, to interact with examinees (Bennett, 1999, Quellmalz & Haertel, 1999). We are called upon to investigate the success of technologies in instruction, even as they target knowledge and skills that are not well measured by conventional assessments.

But how do we design complex assessments so they provide the information we need to achieve their intended purpose? How do we make sense of data of the complex data they may generate? In and of themselves, the rules-of-thumb and the statistical methods that evolved to manage classroom quizzes and standardized tests offer scant help. And ample evidence casts doubt on the strategy of first constructing and administering a complex assessment, then hoping to figure out 'how to score it' only after the data are in.[2]

This presentation is based on two premises: First, that the principles of evidentiary reasoning that underlie familiar assessments are a special case of more general principles. Second, that these principles can help us design and analyze new kinds of assessments, with new kinds of data, to serve new purposes. General discussions of the central ideas can be found in the work of scholars such as David Schum (1987, 1994), who has written extensively on evidentiary reasoning, and Sam Messick (1986, 1992), who has explored assessment design and validity from the perspective of the philosophy of science.

The first half of the paper reviews an 'evidence-centered' framework for designing assessments, developed under Educational Testing Services's PORTAL project (Mislevy, Steinberg, & Almond, in press). The second half discusses, through the lens of this framework, how and where advances in cognitive psychology and technology can be

[2] Don Melnick, who headed the National Board of Medical Examiners' project on computer-based patient-management simulations, recently commented, "It is amazing to me how many complex 'testing' simulation systems have been developed in the last decade, each without a scoring system." (Melnick, 1996, p. 117)

brought to bear to improve assessment. We draw upon three running examples to illustrate ideas throughout the presentation. They are (1) a familiar standardized test, the GRE; (2) a prototype simulation-based assessment of problem-solving in dental hygiene for the Dental Interactive Simulations Corporation (DISC) (Mislevy et al., 1999; in press); and (3) an online performance task, the MashpeeQuest, designed to evaluate students' information analysis skills, as part of Classroom Connect's AmericaQuest instructional program (Penuel & Shear, 2000).

Evidence-Centered Assessment Design

Reasoning from Complex Data

So how should we design and analyze complex assessments? We may begin by asking how people make sense of complex data more generally. Just how do we reason from masses of data of different kinds, fraught with dependencies, hiding redundancies and contradictions, each addressing different strands of a tangled web of interrelationships? It turns out that humans interpret complex data through some underlying "story"—a narrative, perhaps, or an organizing theory, a statistical model, or some combination of these. It might be a simplifying schema we can hold in mind all at once, such as "30 days hath September…", or a complicated structure such as a compendium of blueprints for a skyscraper. We attempt to weave a sensible and defensible story around the specifics. A story that addresses what we really care about, at a higher level of generality and a more basic level of concern than any of the particulars. A story that builds around what we believe to be the generative principles and patterns in the domain.

Just as the principles in the relevant domains are important, so are principles of reasoning from fallible data. Through these latter principles we connect what we know about a domain to what we see in the real world. Over the years, workers in such diverse fields as statistics, jurisprudence, philosophy, and expert systems have discerned some transcending patterns of reasoning from evidence—patterns that are manifest in specialized forms in many domains, each with its own underlying substance and kinds of evidence. We can use these patterns as building blocks when we tackle some novel problem that poses new questions or presents new kinds of evidence. In this spirit, our work in assessment design applies an approach that Schum (1987, 1994) espouses: Structuring arguments from evidence to inference around generating principles in the domain, using probability-based reasoning to manage uncertainty. In

particular, our PORTAL design framework integrates these principles of evidentiary reasoning with principles concerning validity in assessment (Messick, 1986, 1992).

Evidentiary Reasoning and Assessment Design

There are two kinds of building blocks for educational assessment. Substantive building blocks concern the nature of knowledge in the domain of interest, how students learn it, and how they use their knowledge. Evidentiary-reasoning building blocks concern what and how much we learn about students' knowledge from what they say and do. How do we assemble these building blocks into an assessment? The PORTAL project provides an evidence-centered perspective on assessment design, and object definitions and data structures for assessment elements and their interrelationships. In this presentation we draw upon the perspective, and use a high-level description of the central objects and interrelationships. This section sketches out the latter, the PORTAL "conceptual assessment framework" (CAF). We will then use the structure of the CAF to discuss where and how advances in psychology and technology can be put to work.

Figure 1 is a high-level schematic of the CAF, showing three basic models we suggest must be present, and must be coordinated, to achieve a coherent assessment. A quote from Messick (1992, p. 17) serves to introduce them:

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

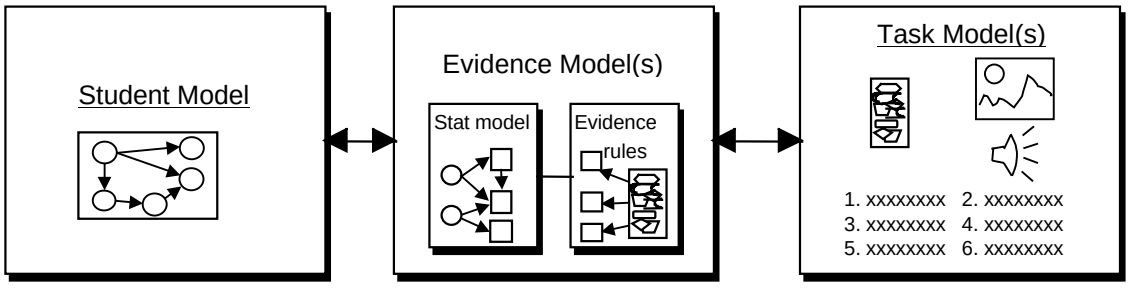


Figure 1. Three Basic Models of Assessment Design.

The Student Model

"What complex of knowledge, skills, or other attributes should be assessed?" This is what the student model is about. Configurations of values of student-model variables are meant to approximate, from some perspective about skill and knowledge in the domain, certain aspects of the infinite configurations of skill and knowledge real students have. It could be the perspective of behaviorist, trait, cognitive, or situative psychology, or some combination of these. As we shall see, this is an area to which advances in psychology have much to contribute. But whichever perspective we wish to weave our stories from, we encounter the evidentiary problem of constructing these stories from limited evidence. Student-model variables are the terms in which we want to talk about students—the level at which we build our story, to determine evaluations, make decisions, or plan instruction—but we don't get to see the values directly. We just see what the students say or do, and must use it as evidence about the student-model variables.

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

A conception of competence in the domain is necessary for determining the number and nature of the student model variables to use in a given application, but it is not sufficient. This critical design decision will also depend on the purpose of the assessment. A single variable that characterizes overall proficiency might suffice in an assessment meant to support only a summary pass/fail decision, for example. But a coached practice system to help students develop the same proficiency—even the same students and the same tasks—would require a finer grained student model, in order to monitor particular aspects of skill and knowledge for which feedback is available.

And when the purpose is program evaluation, the grainsize and the nature of the student model variables should reflect hypothesized ways in which a program may enjoy more or less success, or promote students' learning in some ways as opposed to

others. In our running example on evaluation, for instance, we see that the purpose of the MashpeeQuest assessment is to gather information about students' information-gathering and synthesis skills in a technological environment. It follows immediately that the student model should include variables that concern aspects of these skills, to be defined more concretely by the kinds of observations we will posit constitute evidence about them.

It requires further thought to decide whether to include student-model variables for aspects of these skills as they are used in nontechnological situations, to be informed by observations from nontechnological situations. There are two reasons one might do this, and both revolve around purpose: Just what kind of story must we be able to construct around the variables of the student model if the assessment is to answer the questions we need to address? First, if we want talk about differential impacts in different environments, we must be able to distinguish skills as they are used in different technological environments. This might be done with a multivariate student-model with variables that disentangle such effects from the same complex performances, or with multiple but distinct assessments with different sources of evidence and each with its own student model variables. Second, if we want to compare students in the targeted instructional program with students not in that program, we will not be able to obtain evidence from the latter with ways of collecting evidence that depend on being familiar with technologies specific to the program.

Example 1: The GRE

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

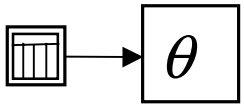


Figure 2. The Student Model for a GRE Measure (Q, V, or A).

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

The GRE's roots are in trait psychology: Verbal ability, quantitative ability, reasoning ability. But "an alternative interpretation of test scores as samples of cognitive processes and contents, and of correlations as indicating the similarity or overlap of this sampling, is equally justifiable… The evidence from cognitive psychology suggests that test performances are comprised of complex assemblies of component information-processing actions that are adapted to task requirements during performance" (Snow & Lohman, 1986). Recognizing the cognitive and social factors that underlie performance on the GRE does not change its predictive value, but it provides a better understanding of just what the GRE tells us and what it doesn't, where it may fall short and for what reasons, when additional evidence may be required and why.

Example 2: DISC

ETS is working with the Chauncey Group International (CGI) to develop a "scoring engine" for a prototype of a simulation-based assessment of problem-solving in dental hygiene, under contract with the Dental Interactive Simulation Corporation (DISC). We are working through student, evidence, and task models with DISC, and consequently examining the implications for the simulator. Two considerations shaped the student model for the prototype assessment. First was the nature of skills DISC wanted to focus on: the problem-solving and decision-making skills a hygienist employs on the job. The second

was the purpose of the assessment: a licensure decision, with some supplementary information about strengths and weaknesses. We will therefore refer to the student model described below as an "overall proficiency + supplementary feedback" student model.

Adapting cognitive task analysis methods from the expertise literature (Ericcson & Smith, 1991), we captured and analyzed protocols from hygienists at different levels of expertise as they solved a range of tasks in the domain (Mislevy, et al., 1999). We abstracted 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. An example was *Using disparate sources of information.* Novice hygienists were usually able to note important cues on particular forms of information, such as shadows on radiographs and bifurcations on probing charts, but they often failed to generate hypotheses that required integrating cues across different forms. We defined student model variables that would characterize a hygienist's tendency to demonstrate these indicators, overall and as broken down into a small number of facets that could also be reported to students. Figure 3 is a slightly simplified version of the model we are presently using.

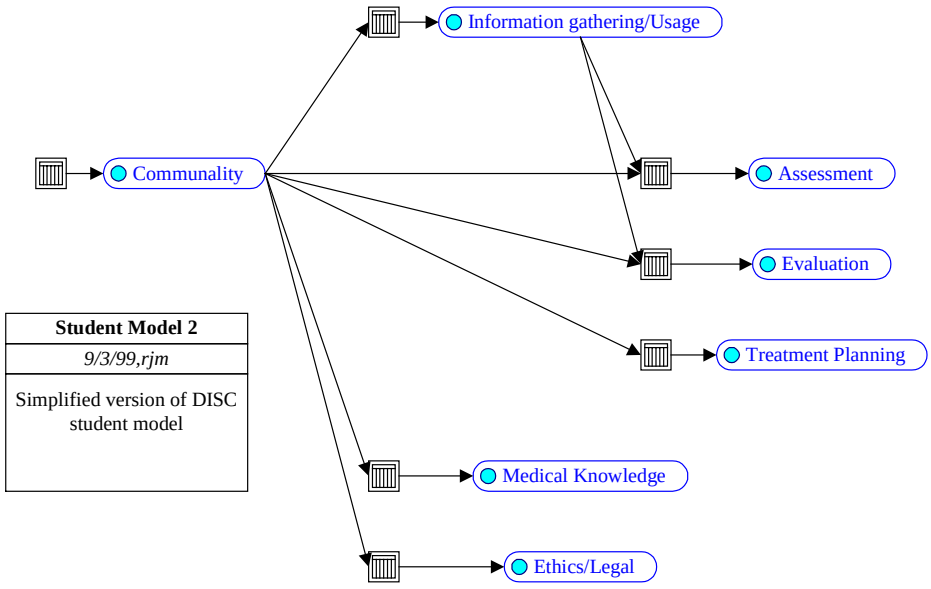


Figure 3. Simplified DISC Student Model.

Figure 3 depicts the Bayes net fragment that contains just the student model variables. As mentioned above, it which can be joined with evidence-model

(EM) fragments that include observable variables so we can to update our beliefs about the student model variables in light of data. The ovals in Figure 3 are the SM variables. Two toward the upper right are *Assessment*, or proficiency in assessing the status of a new patient, and *Information-gathering/Usage*. The full model further elaborates *Information-gathering/Usage*, into variables for knowing how and where to obtain information, being able to generate hypotheses that would guide searches and interpretations, and knowing how to gather information which would help confirm or refute hypotheses.

Example 3: MASHPEEQUEST

Our third running example is a on-line performance task designed by researchers from SRI International, in connection with the evaluation of Classroom Connect's AmericaQuest instructional program. One of the goals of AmericaQuest is to help students learn to develop persuasive arguments, supported by evidence they acquire from the course's Web site or their own research. MashpeeQuest poses a problem that gives students anopportunity to put these skills to use, in a Web-based environment that structures their work.

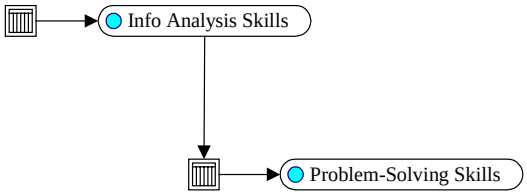
The design of the MashpeeQuest performance task was motivated by the goals of the evaluation. It assesses a subset of the skills that the AmericaQuest program is meant to foster. The skills, broken down into two main categories and constituent subskills, are the following:

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

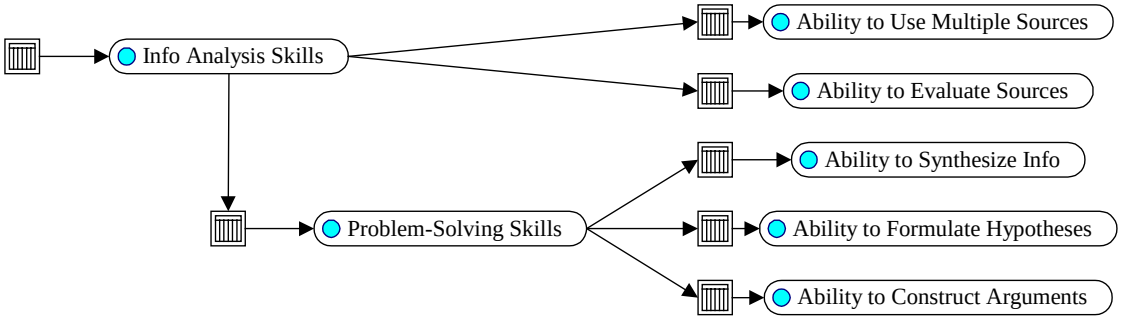
- *Problem Solving Skills.* Ability to synthesize disparate ideas through reasoning in a problem-solving context; ability to offer reasoned arguments rather than brief guesses; ability to formulate creative, well-founded theories for unsolved questions in science and history

Figure 4 illustrates two possible student models that are consistent with the preceding description. They differ in their level of specificity, or grain-size. The first contains just two variables, and would be used to accumulate information about students in terms of just *Information Analysis Skills* and *Problem-Solving Skills*. The arrow between them indicates that they may be correlated in the population of students being addressed. The second student model includes variables for the subskills, so that evidence may be accumulated separately for them, and used to identify for students or teachers more specific areas of strengths or difficulties. Deciding which of the two models to use would require (1) weighing the more detailed information in the finer-grained model against its

lower accuracy, and (2) examining the empirical correlation among the subskills, since the more highly they are correlated the less that is gained by modeling them explicitly.



a) A Coarsely-grained Student Model



b) A Finer-grained Student Model

Figure 4. Possible MashpeeQuest Student Models.

The effective meaning of any of these student model variables will be determined by choices about the observations that are deemed to constitute evidence about them. In the MashpeeQuest task, students will have to weigh evidence they might find in on-line visits to cities in the northeastern United States to help decide a court case involving recognition for the Mashpee Wampanoags, a Native American tribe in Massachusetts. In AmericaQuest, students participate via the Internet in a bicycling expedition with archaeologists and historians who are uncovering clues about the fate of another Native American tribe, the Anasazi, who abandoned their magnificent cliff dwellings in large numbers between 1200 and 1300. Many believe the Mashpee Wampanoags disappeared, just as some people believe the Anasazi disappeared at the time of the abandonment A band of people claiming Wampanoag ancestry have been trying for over twenty years to gain recognition from the federal government as a

tribe that still exists. In 1978, a federal court ruled against the Mashpee Wampanoags' claim, arguing that the tribe could not prove that it had a continuous stake on territory in Mashpee. The tribe is seeking recognition a second time in court. The assessment asks students to take a position on the case, and to identify places where a Quest expedition team should go based on information about the kinds of evidence they might find there. Students will be asked to investigate the evidence, select sites that provide evidence to support their claim, and justify their choices based on the evidence. In addition, they will be asked to identify one place to go to find evidence that doesn't support their claim, and to address how their theory of what happened to the Mashpee Wampanoags is still justified.

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

MashpeeQuest is a single, extended, performance task. In the following discussions, we will think of it as one instance from a class of tasks that might be constructed to obtain evidence in similar ways about the Information Analysis and Problem-solving skills that are being addressed in this particular evaluation project.

Evidence Models

"What behaviors or performances should reveal those constructs," and what is the connection? This is what evidence models are about. The evidence model lays out our argument about why and how the observations in a given task situation constitute evidence about student model variables. Figure 1 shows two parts to the evidence model, the *evaluative* submodel and the *statistical* submodel. The evaluative submodel extracts the salient features of the "work product," that is, whatever the student says,

does, or creates in the task situation. The statistical submodel updates the student model in accordance with the values of these features, effectively synthesizing the evidentiary value of performances over tasks.

The Evaluative Submodel. In the icon for the evaluative submodel in Figure 1, the work product is represented by a rectangle containing a jumble of complicated figures at the far right. It is a unique human production, as simple as a mark on an answer sheet, as complex as the presentation of disconfirming evidence or a series of treatments in a patient-management problem. The three squares coming out of the work product represent "observable variables," the evaluative summaries of what the assessment designer has deemed the key aspects of the performance to take away as nuggets of evidence. The evaluative rules say how to map unique human productions into a common interpretative framework. They embody our beliefs about what is important, for the assessment's purposes, in a performance. These mappings can be as simple as determining whether the mark on an answer sheet is the correct answer, or as complex as an expert's evaluation of multiple aspects of an unconstrained patient-management solution. They can be automatic or they can require human judgment.

Example 1, the GRE, continued

This is an evidence rule in the GRE P&P test:

> IF the response selected by the examine matches the
>
> response marked as the key in the database,
>
> THEN the item response IS correct
>
> ELSE the item response IS NOT correct.

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

Example 2, DISC, continued

The DISC cognitive task analysis (CTA) produced 'performance features' that characterize patterns of behavior and differentiate levels of expertise. They are grist for defining generally-defined, reusable observed variables in evidence models. The evidence models themselves are structured assemblies of student-model variables and observable variables, including methods for determining the values of the observable variables and updating student-model variables accordingly. A particular case will utilize the structures of one or more evidence models, fleshed out in accordance with specifics of that case.

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

Example 3, MashpeeQuest, continued

The MashpeeQuest task requires students to demonstrate a particular set of information analysis skills and problem solving skills that form the student model. While the task provides only a single problem context in which students may demonstrate these skills, it provides multiple opportunities for students to demonstrate different aspects of information analysis skill, in different ways, in different parts of the problem. The observable variables defined to evidence information skills all demonstrate more generally cast skills one needs to use the Internet to conduct research or inquiry: comparing information from multiple sources by browsing and reading different Web links; constructing texts that compare information gleaned from these sources; and evaluating the credibility of that information. The observables evidencing problem-solving skills are specific to the AmericaQuest instructional program, but all have strong parallels to the argumentation skills required of students in

other innovative Web-based learning programs (e.g., Linn, Bell, & Hsi, 1999). These include using information on the Internet as clues to solve a discrete problem, and generating theories, based on consideration of evidence and counterevidence, related to a controversy in history and anthropology.

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

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

Example 1, the GRE, continued

Figure 5 shows the statistical submodel of the evidence model used in the GRE CAT, an item response theory (IRT) model. On the left is a fragment of Bayesian inference network for updating the probability distribution of the student's proficiency parameter given a response to a particular Item *j*. The distribution between the student-model proficiency variable θ and the item response X_j represents the conditional probability distribution of X_j given θ. The probability distribution for θ, which appears in Figure 2, is not part of the evidence model. The evidence model contains the functional form of the item response model, the previously-estimated item parameters, and the indication of which student-model parents are posited to produce responses—in this just, since it is a univariate IRT model, simply a pointer to θ. When it is time to make an observation, this fragment is "docked" with the SM BIN fragment to form a complete probability model for θ and X_j jointly. On the right of the figure is a library of all items that could be given, along with the structures necessary to dock any one with the student model in order to incorporate the evidence its response contributes. Further discussion of the statistical aspects of this process can be found in Mislevy, Almond, Yan, & Steinberg (1999). The information stored along with these fragments also provides the data needed for selecting the next item to administer to optimize information about an examinee's θ. Assembling statistical models as they are needed from a collection of prefabricated fragments is an example of what has been called 'knowledge-based model construction' in the expert systems literature (Breese et al., 1994).

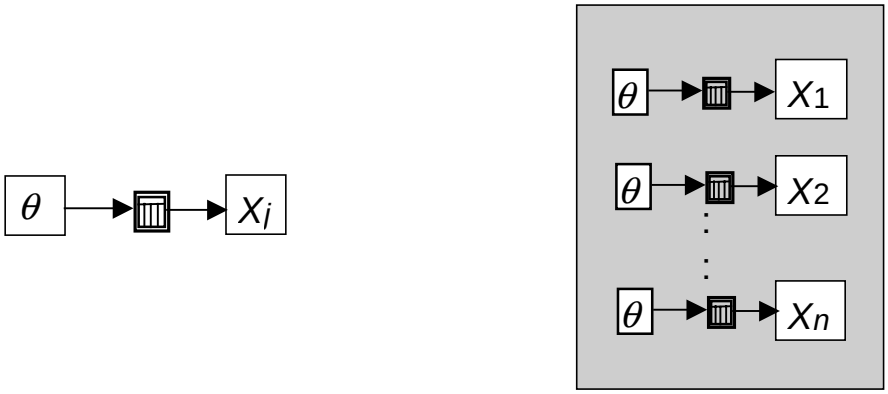


Figure 5. The Statistical Submodel of the Evidence Model in the GRE-CAT.

Example 2, DISC, continued

Figure 6 shows the Bayes net fragment that comprises the statistical submodel of one particular evidence model. It concerns gathering patient information when assessing a new patient's status. At the far left are student model variables we posit drive performance in these situations: *Assessment* of new patients and *Information-gathering /Usage*. The nodes on the right are generally-defined observable variables. Two of them are *adapting to situational constraints*, and *adequacy of examination procedures,* in terms of how well their rationale is grounded. *Adequacy of examination procedures,* for example, has three values: *All* of the necessary points of an appropriate rationale are present in the examinee's solution; *some* are present; or *none* or few are present. These are generic categories that will be particularized for actual specific cases, the part of the evidentiary argument that is addressed in the evaluation submodel discussed above. In a specific case, then, the values of the observable variables are the result of applying rubrics—in this assessment, computer algorithms applied to the examinee's sequence of actions. Figure 7 shows how this evidence model Bayes net fragment is 'docked' with the student model fragment when an examinee is working in a situation that has been constructed to conform with this evidence model.

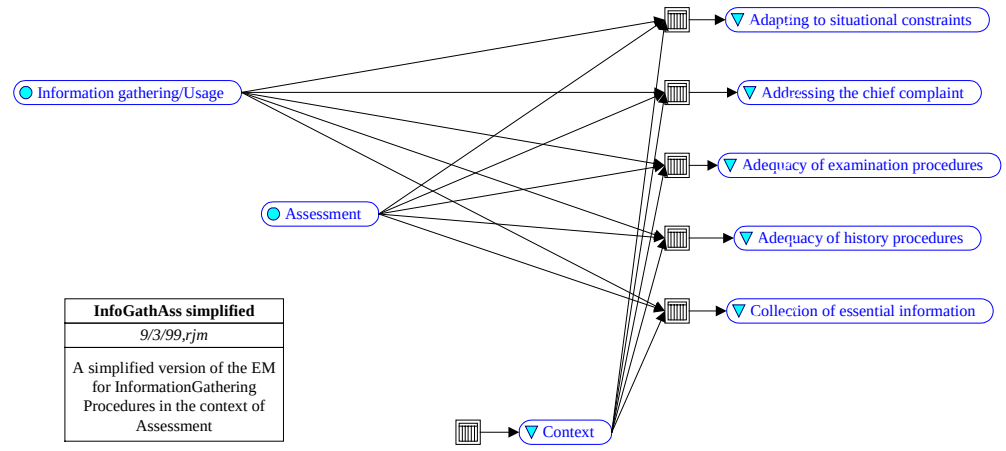


Figure 6. The Bayes Net Fragment in an Evidence Model in DISC.

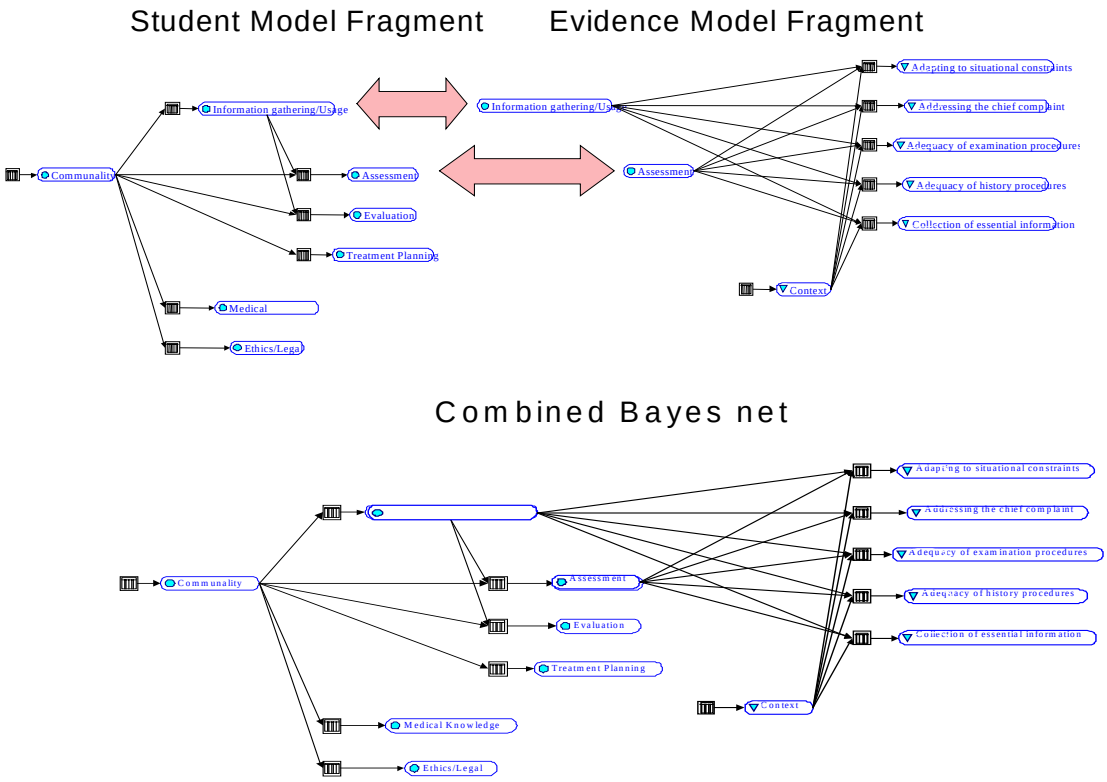


Figure 7. Docking Student Model and Evidence-Model Bayes Net Fragments in DISC.

Example 3, MashpeeQuest, continued

Figure 8 depicts the statistical submodel of the evidence model related to student information analysis skills assessed in a hypothetical family of tasks like the MashpeeQuest task. The focus is on measuring student performance in the context of a problem that requires them to read, interpret, and use information on the Web to solve a problem like those presented in the AmericaQuest program. At the left of the figure are two variables from the finer-grained student model introduced above, namely the *Ability to use multiple sources* and *Ability to evaluate sources* that are parts of *Information analysis skills*. These parent variable drive the probabilities of the observable variables in the middle of the figure and the lower right. We see that *Ability to use multiple sources* is informed by the observable variables *Number of sources*, *Time per source*, and [quality of] *Comparison across links*. *Number of sources* could have as many values as are links in the task. Because no prior information is given to students about what sources are more likely to have useful information, more sources considered is taken as evidence of better information analysis skills. *Time per source* could have any number of values from just a few seconds to several minutes. Here, one would see if students

were simply "clicking through" without reading a particular link. The time spent is an important counter-balance to the number of sources considered, since it is an (albeit imperfect) indicator of whether students actually read the text on the links they used. [Quality of] *Comparison across links* is actually a composite of two ratings of the same student responses, namely evaluations of how well they discussed the *Coherence* and the *Credibility* of the sites they visited—key features of effective information analysis, according to experts in this domain (Wineburg, 1998).

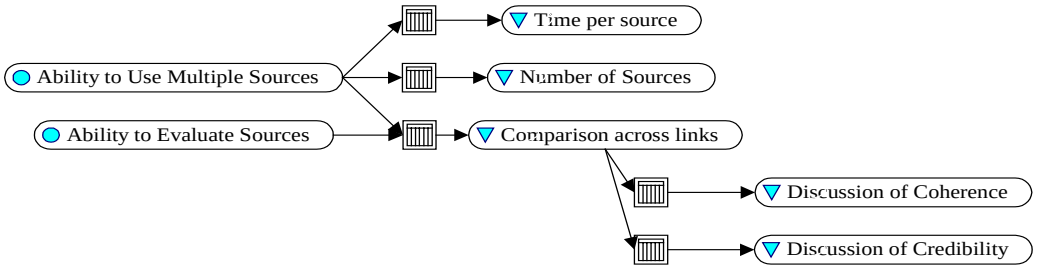


Figure 8. A MashpeeQuest Evidence Model.

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

Task Models

"What tasks or situations should elicit those behaviors?" This is what task models are about. A task model provides a framework for constructing and describing the situations in which examinees act. A task model includes specifications for the

environment in which the student will say, do, or produce something; for example, characteristics of stimulus material, instructions, help, tools, affordances. It also includes specifications for the work product, the form in which what the student says, does, or produces will be captured.

Assigning specific values to task model variables, and providing materials that suit the specifications there given, produces a particular task. A task thus describes particular circumstances meant to provide the examinee an opportunity to act in ways that produce information about what they know or can do more generally. The task itself does not describe what we should attend to in the resulting performance or how we should evaluate what we see. This is specified in the evidence model. Distinct, possibly quite different, evidence rules could be applied to the same work product from a given task. Distinct and possibly quite different student models, befitting different purposes or derived from different conceptualizations of proficiency, could be informed by data from the same task.

Developments in cognitive psychology have much to say about how we might construct situations that can provide evidence about what students know and can do. Developments in technology provide the means for producing these situations, and capturing, evaluating, and communicating what students do there.

Example 1, the GRE, continued

A task model in the GRE describes a class of test items. There is some correspondence between task models and GRE "item types" (e.g., sentence completion, passage comprehension, quantitative comparison). Different item types will generally require different task models, because different sets of variables needed to describe their distinct kinds of stimulus materials and presentation formats, and different features may be important in modeling item parameters or controlling item selection. Different task models will be required for P&P and CAT use of what is the same item from the perspective of content, because specifications for presenting and managing the item are wholly different in the two modes.[3]

[3] The item types of the GRE trace their roots back to the Army Alpha intelligence test that Robert Yerkes administered to millions of American recruits in World War I—pure measures of verbal and quantitative reasoning ability, it was thought. Some of the original tasks have fallen by the wayside over the years, recognized as requiring knowledge rooted too deeply in the incidentals of class and culture. The survivors hold up surprisingly well under contemporary psycholinguistic and cognitive analyses. As Snow and Lohmann suggest, most of them can indeed be analyzed in terms of knowledge of skills and

Example 2, DISC, continued

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

There are groups of task model variables, and the variables in some of them are hierarchical. For example, *Oral Hygiene Status* is a TM variable with possible values {excellent, good, poor, extremely poor}. For example, *Documentation Status* includes *Documentation Age*, *Documentation Completeness*, and *Documentation Availability.* The documentation status variables are important for evoking evidence about particular aspects of examinees' proficiencies in obtaining information. If we want to learn about an examinee's ability to seek and interpret information that is not initially present—and the CTA told us this is indeed an important aspect of competence in dental hygiene—then we cannot have a case in which *Documentation Completeness* is set at "all information presented." The interactive capabilities of the computer-based simulation make it possible to gather evidence about this aspect of proficiency.

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

Task models also include specifications for work products. The simulator records the sequence of actions an examinee makes, which can then be parsed by evaluation rules. The CTA also suggested the value of some more structured work products. Hygienists at different levels of expertise were distinguished not only by the actions they took, but the reasons they gave for taking them. Several performance features concerned intermediate mental products such as identification of cues, generation of hypotheses, and selection

---

procedures one needs to gain and use information from prose or from mathematical settings, and prove hard or easy for the right reasons.

of tests to explore conjectures—steps that are usually not manifest in practice, but which directly involve central knowledge and skills for problem-solving in dental hygiene. Work products that require the examinee to make such steps explicit will capture more direct evidence of the thinking behind a solution than the sequence of actions will. Following patient assessment, for example, the examinee will fill out a summary form that requires synthesized findings in a form similar to commonly used insurance forms. From the perspective of situated psychology, knowing how to use forms like this is an integral part of expertise in the dental care community.

Example 3, MashpeeQuest, continued

In the assessment designed for AmericaQuest, a number of features of the task model would be present, whether or not a task addressed the specific problem of explaining the continued existence or disappearance of the Mashpee Wampanoags. The most critical features of the motivating problem context are not content specific, even though they are subject-matter specific. The kinds of problems should involve the consideration of historical and archaeological evidence as AmericaQuest does, which does not necessitate a focus on the Mashpee or any other Native American tribe per se. The problem statement should ask students to formulate a hypothesis and back it with evidence gathered from information available to them in the Web-based assessment environment, as they do in AmericaQuest and as specified in the student model of problem solving skill. The task model would vary if students were asked to display analysis skills in ways other than stating a hypothesis and supplying Web-based evidence, for then both the kinds of material made available to students and the kinds of work products they produced could differ. For example, students might be asked to use a particular digital library, which contained a vast collection of materials, to demonstrate their analytic skills, Not only would they specify a hypothesis and provide supporting evidence, they might also produce an intermediate work product that documents: (1) the searches they conducted of the library's relational database; (2) the path they navigated as they accessed the library's various collections, (3) the different search and retrieval strategies they used to gather evidence; and (4) the ways they could state and post questions in a discussion forum to gather additional information or help. These intermediate work products can be used to make students' thinking more explicit.

The task-specific observable variables in the evidence model for an alternative task would vary with content, but would maintain the same critical components. For example, consideration of multiple sources of information

that advance competing hypotheses about the problem and evaluation of those sources is critical. If students can see all the information on a single page produced by a single author, there are no grounds for arguing for or against a particular hypothesis, and students' analytic skills are only minimally engaged compared to basic text comprehension skills. The key sources of evidence of student skill in analyzing information is whether students consider evidence from sources that support competing theories about the historical dilemma at hand. The particulars of the problem itself matter less. For example, one could imagine another task in which students are given several Internet links to visit related to the controversy over whether Marco Polo ever traveled to China, and are then asked to evaluate the sites for their relevance to solving this historical dilemma. The central features of the task model are that the problem involve historical subject matter focused around a controversy or competing hypotheses and that students be required to review multiple sources with evidence that supports or disconfirms particular hypotheses.

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

Leverage Points for Improving Assessment

This has been a quick tour of a schema for the evidentiary-reasoning foundation of assessments. It gives us some language and concepts for talking about this central core of assessment, not only for familiar forms and uses of assessment, but for new forms

and uses. We can use this framework to discuss ways we can take advantage of advances in psychology and technology. Along the way, we also point out ways that developments in statistics can play a central role in bringing the promise to fruition.

Leverage Points for Psychology

While the familiar practices of assessment and test theory originated under the regimes of trait and behaviorist psychology, contemporary views of learning and cognition fit more comfortably into the headings of cognitive and situative psychology (Greeno, Collins, & Resnick, 1997). The cognitive perspective includes both the constructivist tradition originated by Piaget, and the information-processing tradition developed by Newell and Simon, Chomsky, and others. The focus is on patterns and procedures individuals use to acquire knowledge and put it to work. The situative perspective focuses on the ways individuals interact with other people in social and technological systems, so that learning includes becoming attuned to the constraints and affordances of these systems. In this paper, we use the term "cognitive psychology" broadly to encompass both of these perspectives.

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

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

> At various stages of learning, there exist different integrations of knowledge, different degrees of procedural skill, differences in rapid access to memory and in representations of

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

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

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

> *Knowing*, in [the situative] perspective, is both an attribute of groups that carry out cooperative activities and an attribute of individuals who participate in the communities of which they are members. … *Learning* by a group or individual involves becoming attuned to constraints and affordances of material and social systems with which they interact. (Greeno, Collins, & Resnick, 1997, p. 17)

These insights challenge the familiar strategy of assessing through standardization—'measuring the same thing' for all students by gathering the same data under the same conditions. For example, AmericaQuest is intended to develop

student skill in analyzing information and problem solving specifically in the context of an Internet-based adventure learning experience. The adventure involves using inquiry tools and evidentiary reasoning skills typically used by historians and archaeologists, but in an important sense, the analysis and problem solving skills students are learning are confounded with learning how to use the Internet to conduct inquiry. Observation data suggests that teachers off-line instruction mediates students' learning these skills, however, in significant ways (Shear & Penuel, 2000). If teachers' assignments to students are unrelated to the central historical dilemma posed by the Quest and students are not directed to weigh evidence about particular hypotheses, students will fail to learn (at least through AmericaQuest) the information analysis and problem solving skills identified in the student model.

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

What we can observe to give us evidence. Given the terms in which we want to characterize students' capabilities, what can we observe that will constitute evidence of

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

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

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

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

The kinds of historical reasoning behaviors elicited in the MashpeeQuest example are behaviors that are parallel to the activities of professional historians. Expert historians spend much of their time analyzing historical texts, images, and artifacts (Wineburg, 1991), just as students in the MashpeeQuest task spend most of their time reading and interpreting the text on the various links to cities in the task. The MashpeeQuest scoring rubric would assign higher scores to student behaviors that

suggested that students were not just spending time analyzing documents but that they were also analyzing them in ways that are similar to the ways expert historians analyze documents (see Wineburg, 1998). Expert historians, for example, may consider how evidence in one document supports or contradicts evidence in another document, something that students are explicitly invited to consider in the MashpeeQuest task. Student skill in analyzing documents is made visible through the formulation of an argument backed by specific evidence from the documents, as well as a consideration of possible counter-evidence from other links on the MashpeeQuest site.

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

Effective ways to elicit the kinds of behavior we need to see. What characteristics of problems stimulate students to employ various aspects of their knowledge? We are beginning to hear phrases such as 'principled task design' in assessment more often nowadays (e.g., Embretson, 1999). The idea is that by systematically manipulating the features of task settings—that is, controlling the constraints and the affordances—we create situations that encourage students to exercise targeted aspects of skill and knowledge. We describe these features in terms of task model variables.

Work on systematic and theory-based task design dates back at least half a century. We may point to Louis Guttman's (1959) facet design for tests, followed by Osburn's (1968) and Hively, Patterson, and Page's (1968) work in the 60's with item forms, and John Bormuth's (1970) linguistic transformations of texts to produce comprehension items. But now we can take advantage of concepts and methods from psychology to build tasks more efficiently and around cognitively relevant—and

therefore construct relevant—features. We have discussed ways we can manipulate the medical conditions of patients and the availability of information in the DISC simulator environment, to either elicit evidence about hygienists' medical or information-gathering knowledge, or to minimize the stress on this knowledge in order to highlight other aspects of their competence. We have also explored how a web-based environment can be used to student information analysis and problem solving skills across a range of tasks; in particular, we have seen considered how the content available at different Internet links could be varied to isolate particular information analysis and problem solving skills A web-based environment is a particularly adaptable vehicle for presenting assessment tasks. The wealth of information available on the Web makes it possible to vary the substance of the assessment task relatively easy, within an assessment schema under which task format and underlying targeted skills remain constant.

Leverage Points for Statistics

This section is a bit of an aside, since the focus of the paper is on ways cognitive psychology and technology can improve assessment. But making sense of data from complex tasks, sorting out its import for the several interrelated aspects of competence in a complex student model, is a difficult challenge. We have mentioned that we use Bayesian inference networks, or Bayes nets, to accomplish this task when we must model the relationships formally, and familiar, off-the-shelf models from test theory are not up to the job. We could not do this as recently as ten or fifteen years ago. It has become possible only through advances in the field of statistics. This section notes some of the most important ones for assessment, and shows where in the CAF they play their roles.

Managing uncertainty with respect to the student model. One way of capitalizing on advances in statistical theory is using Bayesian inference networks, or Bayes nets for short, to manage our knowledge about student model variables. The basic idea behind Bayes nets is to model interrelationships among large numbers of variables in terms of local interactions within small subsets, using conditional independence structures. We thus synthesize the structuring of inference from building-block relationships and the belief calculus of probability (Schum, 1994). The structure of a Bayes net corresponds to a part of a validity argument. Conditional probabilities within the net are the basis of extensions of reliability and precision arguments. We can express test theory in this framework, and use the same framework

to help extend the generative principles behind test theory to more complex problems. Complementary developments in estimation methods, notably Markov Chain Monte Carlo methods (Gelman et al., 1995), allow us to build and fit models from building block distributions very flexibly.

The work of VanLehn and his associates (e.g., Martin & VanLehn, 1995; Schulze et al., 2000) on modeling problem-solving in physics illustrates a synergy among technology, statistics, and cognitive psychology, with principled reasoning made possible only through advances in statistics. VanLehn models undergraduate students' understanding of mechanics at a fine grain size in the OLEA and ANDES intelligent tutoring systems (ITSs). From cognitive psychology, he builds a student model around production rules—condition-action relationships that comprise much of problem-solving ability in this domain. From technology, he allows students to work through problems step by step, with a computer keeping track of their problem steps, modifying the problem space as they build a solution, parsing and evaluating each move. From statistics, he assembles Bayes nets on the fly to model a student's solution through a given problem, then collapses the information after the problem into summary variables that determine prior distributions for the finely-grained model of the next problem. In sum, VanLehn's approach illustrates several of the roles that technology can play in assessment—to present, monitor, interact, customize, and document an assessment and learning task to the student's needs. In particular, technology can yield rich evidence about a student's reasoning processes.

Managing the relationship between observations and student model variables. While the student model in the GRE couldn't be simpler, the concept of computerized adaptive testing (CAT) is quite sophisticated (Wainer et al., 1990). Different items are presented to different examinees, which makes it tough enough to bring everyone's performance into the same summary metric. More challenging is the fact that the sets of items different people take are not of the same difficulty. Even more challenging, they are selected one at a time to be optimally informative about the student, in light of what we have already learned from her preceding responses. The success of CAT depends on the modular structure of the chains of reasoning we construct from abilities to response probabilities to item responses, and back again to an updated ability distribution via Bayes theorem once we observe a response. As Figure 5 showed, this reasoning and a pool of items with item parameters, is the statistical underpinning of CAT.

The DISC project and the ANDES ITS show that we can apply the same strategy with more complex tasks (Figure 9). We can have a library of these evidence model Bayes net fragments, and dock them with the student model on the fly when observations are obtained for the task at hand. As we stressed in the DISC example, the key to success is to recognize, through the lens of cognitive psychology, recurring patterns in problem situations around which we can build re-usable statistical model fragments. These evidence model Bayes net fragments embody the structure of the evidentiary argument, to be tailored for specific problems by instantiating the values of task model variables and in turn conditional probabilities.

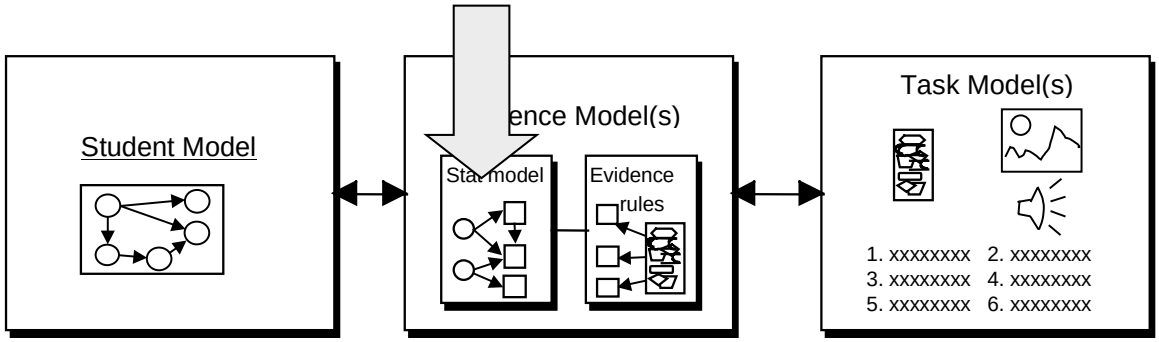


Figure 9. Managing the Evidentiary Relationships Between Observations and Student-Model Variables.

Extracting features and determining values of observable variables. Statistical methods can also be useful in the stage of extracting and evaluating values of observable variables from work product—that is, the evaluation submodel of the CAF evidence model (Figure 10). One way is to use Bayes nets as part of the feature evaluation; i.e., using Bayes nets for pattern recognition, to classify the unique work product as a likelihood over predetermined possibilities of a possibly vector-valued "observable" variable. For this job, alternatives include neural networks (Stevens et al., 1996), rule-based evaluations (Clauser et al., 1997), fuzzy logic (Hawkes & Derry, 1989 / 1990), and human ratings. The second way statistical methods can help here is to model ratings by human judges: detecting outliers, calibrating their severities, and characterizing the uncertainty associated with different rating configurations (Linacre, 1989; Patz & Junker, 1999).

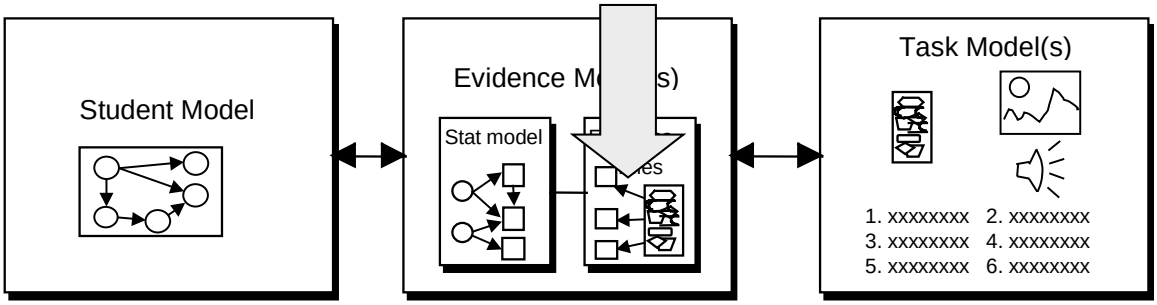


Figure 10. Extracting and Evaluating Features of Complex Performances.

Leverage Points for Technology

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

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

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

variables that answer a question like, "what is the *nature* of Meredith's proficiency, now that we know the *level* of her proficiency?"

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

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

Automated extraction and evaluation of key features of complex work. Some automated extraction and evaluation of key features of complex work make it possible to increase the efficiency of applying existing evaluation rules, and others make it possible to evaluate work products we could not routinely include in assessment at all

(Figure 10). We have already mentioned that Stevens uses neural networks to summarize the import of students' sequences of diagnostic tests. Examples from current projects at ETS include:

- natural language processing methods for scoring essays, with psycholinguistic and semantic theory to define features to extract and tree-based regression to summarize them into scores;

- evaluation of constructed show-your-steps, responses to algebra problems, with GOMS methodology to infer students' likely strategies; and

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

Examples from MashpeeQuest include:

- counts of the number of Internet links checked and calculation of the amount of time spent examining each link

- evaluation of student reasoning by identifying whether evidence from particular links are used to support particular hypotheses

- comparison of students' own ratings of the relevance of particular links' with experts' ratings.

Automated/assisted task construction, presentation, and management. A preceding section discussed how research in cognitive psychology reveals systematic relationships between, on one hand, the affordances and constraints of problem situations, and knowledge structures and procedures people can bring to bear on those problems. Understanding and systematically manipulating these features of tasks not only helps us produce tasks more efficiently, it strengthens the validity argument for them as well. Further benefits accrue if we can use technology to produce tasks as well. This is as true for producing familiar kinds of tasks as it is for ones that could not be exist at all outside a technological setting (such as DISC's computer-based simulations). Likewise, the VideoDiscovery technology-based investigations and the SMART assessments developed by the Cognition and Technology Group at Vanderbilt illustrate the use of technology to assess phenomena that are too large, too small, too dangerous, too dynamic, too complex, or too dangerous to be validly assessed using non-

technology based methods of assessment (Vye et al., 1998.) The production side of assessment can exploit technology in several ways, including, for example, automated and semi-automated construction of items (e.g., Bennett, 1999) and tools to create tasks according to cognitively-motivated schemas (e.g., Embretson, 1998).

A further comment on technology-based assessment. Technology is as seductive as it is powerful. It is easy to spend all one's time and money designing realistic scenarios and gathering complex data, and only then ask "how do we score it?" When this happens, the chances are great that the technology is not being used to best effect. The affordances and constraints are not selected optimally to focus attention on the skills and knowledge we care about, and to minimize the impact of incidental skills and knowledge. This is why we emphasize the evidentiary foundation that must be laid if we are to make sense of any complex assessment data. The central issues are construct definition, forms of evidence, and situations that can provide evidence, regardless of the means by which data are to be gathered and evaluated. Technology provides such possibilities as simulation-based scenarios, but evidentiary considerations should shape the thousands of implementation decisions that arise in designing a technology-based assessment. These are the issues that cause such an assessment to succeed or to fail in serving its intended purpose. Messick's (1992) discussion on designing performance assessments is mandatory reading for anyone who wants to design a complex assessment, including computer-based simulations, portfolio assessments, and performance tasks.

Example 2, DISC, continued

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

- present the distinct phases in the patient interaction cycle (assessment, treatment planning, treatment implementation, and evaluation);

- present the forms of information that are typically used, and control their availability and accessibility, so we can learn about examinees' information-gathering skills;

- manage cross time cases, not just single visits, so we can get evidence about examinees' capabilities to evaluate information over time; and

- vary the virtual patient's state dynamically, so we can learn about examinees' ability to evaluate the outcomes of treatments that she chooses.

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

- seek and gather data;

- indicate hypotheses;

- justify hypotheses with respect to cues;

- justify actions with respect to hypotheses.

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

Example 3, MashpeeQuest, continued

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

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

- present students with an historical or archaeological dilemma with competing hypotheses to consider;

- present students with distinct phases of problem-solving using historical documentation;

- vary the problem or dilemma, to provide evidence for generalizability of student skills across tasks;

- include multiple sources of pictorial and text-based evidence that can be used tosupport or to disconfirm different hypotheses;

- allow for students to enter a text-based argument regarding their own position about the dilemma;

- vary the outcomes of the search dynamically, so we can learn about students' ability to evaluate the outcomes of searches that she conducts.

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

- seek and gather data on the Internet, or a controlled facsimile;

- carry out analses of the evidence found on as many links as possible in the task;

- construct a coherent argument in support of one hypothesis using evidence from the links, with both confirming and disconfirming evidence that can be discovered and taken into account;

- enter text in a setting that for which Internet etiquette and commonly-used protocol are appropriate.

CONCLUSION

The advancements in technology and psychology discussed here will have the most impact when assessments are built for well-defined purposes, and connected with a conception of knowledge in the targeted domain. They will have much less impact for 'drop-in-from-the-sky' large-scale assessments like NAEP. They are important in two ways for gauging students' progress and evaluating the effectiveness of educational programs.

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

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

REFERENCES

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

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

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

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

Breese, J.S., Goldman, R.P., & Wellman, M.P. (1994). Introduction to the special section on knowledge-based construction of probabilistic and decision models. *IEEE Transactions on Systems, Man, and Cybernetics, 24*, 1577-1579.

Clauser, B.E., Ross, L.P., Clyman, S.G., Rose, K.M., Margolis, M.J., Nungester, R.J., Piemme, T.E., Chang, L., El-Bayoumi, G., Malakoff, G.L., & Pincetl, P.S. (1997). Development of a scoring algorithm to replace expert rating for scoring a complex performance-based assessment. *Applied Measurement in Education, 10,* 345-358.

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

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

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

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

Gelman, A., Carlin, J.B., Stern, H.S., & Rubin, D.B. (1995). *Bayesian data analysis.* London: Chapman & Hall.

Glaser, R., Lesgold, A., & Lajoie, S. (1987). Toward a cognitive theory for the measurement of achievement. In R. Ronning, J. Glover, J.C. Conoley, & J. Witt (Eds.), *The influence of cognitive psychology on testing and measurement: The Buros-*

Nebraska Symposium on measurement and testing (Vol. 3) (pp. 41-85). Hillsdale, NJ: Erlbaum.

Greeno, J.G., Collins, A.M., & Resnick, L.B. (1997). Cognition and learning. In D. Berliner & R. Calfee (Eds.), Handbook of educational psychology (pp. 15-47). New York: Simon & Schuster Macmillan.

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

Hawkes, L. W & Derry, S. J. (1989/90). Error diagnosis and fuzzy reasoning techniques for intelligent tutoring systems. *Journal of AI in Education, 1,* pp. 43-56.

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

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

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

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

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

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

Linacre, J. M. (1989). *Multi-faceted Rasch measurement.* Chicago: MESA Press.

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

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

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

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

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

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

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

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

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

Patz, R.J., & Junker, B.W. (1999). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics, 24,*342-366.

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

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

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

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

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

Schulze, K.G., Shelby, R.N., Treacy, D.J., Wintersgill, M.C. (2000). Andes: A Coached Learning Environment for Classical Newtonian Physics. To appear in *Proceedings of the 11th International Conference on College Teaching and Learning.* Jacksonville, FL, April, 2000.

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

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

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

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

Wainer, H., Dorans, N.J., Flaugher, R., Green, B.F., Mislevy, R.J., Steinberg, L., & Thissen, D. (1990). *Computerized adaptive testing: A primer.* Hillsdale, NJ: Lawrence Erlbaum Associates.

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

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

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