The Role of Probability-Based Inference in an Intelligent Tutoring System[1]

Robert J. Mislevy & Drew H. Gitomer

Educational Testing Service

February, 1995

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

The Role of Probability-Based Inference in an Intelligent Tutoring System

Robert J. Mislevy & Drew H. Gitomer

Educational Testing Service

Abstract

Pursuit of efficient probability-based inference in complex networks of interdependent variables is an active topic in current statistical research, spurred by such diverse applications as forecasting, pedigree analysis, troubleshooting, and medical diagnosis. This paper concerns the potential role of Bayesian inference networks for updating student models in intelligent tutoring systems (ITSs). Basic concepts and tools of the approach are reviewed, but emphasis is on special considerations that arise in the ITS context. We explore how this approach can support generalized claims about aspects of student proficiency through the combination of detailed epistemic analysis of particular actions within a system with probability-based inference. The psychology of learning in the domain and the instructional approach are seen to play crucial roles. Ideas are illustrated with HYDRIVE, an ITS for learning to troubleshoot an aircraft hydraulics system.

Key words: Bayesian inference networks, cognitive diagnosis, HYDRIVE, intelligent tutoring systems, probability-based inference, student models

Contents

Overview

All intelligent tutoring systems (ITSs) are predicated on some form of student modeling to guide tutor behavior. Inferences about a student's current skills, knowledge, and strategy usage can affect the presentation and pacing of problems, quality of feedback and instruction, and determination of when a student has completed some set of tutorial objectives. But we cannot directly observe what a student does and does not *know*; this we must infer, imperfectly, from what a student does and does not *do*. This paper discusses an integration of principles of cognitive diagnosis and principles of probability-based inference, forged in an attempt to develop a generalizable framework for student modeling in intelligent tutoring systems.

Central to the development is the conception of the student model. In any particular application, we work with a set of aspects of skill and knowledge that are important in that application. These are the variables in a space of "student models," particular configurations of values which approximate the multifarious skill and knowledge configurations of real students. Depending on the purpose, one might distinguish from one to hundreds of aspects of competence in a student model space. They might be expressed in terms of categories, qualitative descriptors, numbers, or some mixture of these; they might be conceived as persisting over long periods of time, or apt to change at the next problem-step. They might concern tendencies in behavior, conceptions of phenomena, available strategies, or levels of aspects of developing expertise. The particular form of the student model space in a given application is driven by a conception of the nature and acquisition of competence in the target domain, and the goals and philosophy of the instructional component of the system.

A student model in an ITS can fulfill at least three functions. First, given a set of instructional options, a student model can provide information to suggest which of the available choices is most appropriate for an individual at a given point in time (Ohlsson, 1987). To the degree that an ITS explicitly represents a domain of knowledge and task performance, it should be possible to design instruction at a level of cognitive complexity that facilitates successful performance and understanding (Kieras, 1988). Second, a student model in an ITS enables prediction of the actions a student will take, given the characteristics of a particular problem state and what the system infers about the student's understanding (Ohlsson, 1987). With some understanding of students and problems, one ought to be able to more accurately predict performance than if no model has been specified. The extent to which student actions conform to these predictions is an indication of the validity of the inferences about students made through the student model. Third, the student model enables the ITS to make claims about the

competency of an individual with respect to various problem-solving abilities. These claims can be viewed as data summaries that aid in determining whether a person is likely to negotiate successfully a particular situation or to help the tutor make decisions about problem selection and exit criteria from a program of instruction.

Obviously any student model oversimplifies the reality of cognition (whatever that may be!). In applications, as Greeno (1976, p. 133) points out, "It may not be critical to distinguish between models differing in processing details if the details lack important implications for quality of student performance in instructional situations, or the ability of students to progress to further stages of knowledge and understanding." The nature and the grain-size of a student model in an ITS ought therefore to be targeted to the instructional options available.[1] A model will first need to include cognitive features related to developing performance, as revealed in part by analyses of the skills and understandings needed for accomplished performance. But because accomplished performance derives from the complex structuring of knowledge and skills, a model of student performance in an ITS will also need to represent the interrelationships of target skills and understandings.

In practice, an ITS must work with specific actions that students take in specific situations. The student model mediates between this level of unique and unrepeatable observations, and the higher level of abstraction at which theory about the development of competence and the design of instruction take place. The inferential task consists of (1) establishing a framework for interpreting specific actions in terms suited to guide instruction and (2) characterizing the information these actions convey about variables in the student model. We consider the use of probability-based reasoning as a means for structuring the inferential task. A distinctive feature of the approach is the differentiation between a model for a student's knowledge (i.e., values of variables in a student model space that encompasses key aspects of knowledge) and a model for an observer's state of knowledge about this student model (Mislevy, 1994a).

---

[1] Similar conclusions have been reached regarding expert systems for medical diagnosis. Weiss (1974), discussing the CASNET system, said that "...the resolution of states should be maintained only at a level consistent with the decision-making goal. A state network can be thought of as a streamlined model of disease that unifies several important concepts and guides us in our goal of diagnosis. It is not meant as a complete model of disease" (p. 34).

Probability theory provides powerful mechanisms for explicating relationships, reasoning bi-directionally, criticizing and improving models, and handling evidentiary subtleties—as a consequence of, and, at the same time, at the cost of, constructing a joint distribution of variables whose modeled interrelationships are taken to approximate beliefs about the status and interrelationships of aspects of students' competences and actions. Due to the recent developments we sketch below, this requirement is not as constraining as is often believed. Discussions of the advantages and disadvantages of the probabilistic approach vis a vis alternatives such as fuzzy logic, belief theory, and rule-based reasoning (with more detail and passion than we can muster here), are found in Cheeseman (1985, 1986), Lindley (1987), Pearl (1988), Schum (1979, 1994), and Spiegelhalter (1987, 1989). That the potential of probability-based reasoning for ITSs (and for expert systems in general) cannot be dismissed out of hand, however, only sets the stage for more hard work: to investigate the scope and the limitations of the learning domains, student models, and instructional approaches for which probability-based reasoning can play a viable role, and to develop practicable procedures for realizing its potential.

To this end, this paper discusses the implementation of probability-based reasoning in the HYDRIVE tutoring/assessment system for developing troubleshooting skills for the F-15 aircraft's hydraulics systems (Gitomer, Steinberg, & Mislevy, 1995). In the course of implementing principles of cognitive diagnosis, HYDRIVE uses a Bayesian inference network to express and update student-model variables—even as deductive logic and rule-based inference play complementary roles in the system. Our objective is not to argue dogmatically for the exclusive use of Bayesian probability-based inference in ITS's, but to share our experiences to date in exploring the ways that this well-stocked and time-tested armamentarium of conceptual and practical tools can be gainfully used in this context.

An Introduction to HYDRIVE

The hydraulics systems of the F-15 aircraft are involved in the operation of flight controls, landing gear, the canopy, the jet fuel starter, and aerial refueling. HYDRIVE is designed to simulate many of the important cognitive and contextual features of troubleshooting on the flightline. A problem starts with a video sequence in which a pilot, who is about to take off or has just landed, describes some aircraft malfunction to the hydraulics technician; for example, the rudders do not move during pre-flight checks. HYDRIVE's interface then offers the student several options, such as the following: performing troubleshooting procedures by accessing video images of aircraft components and acting on those components; reviewing on-

line technical support materials, including hierarchically organized schematic diagrams; and making their own instructional selections at any time during troubleshooting, in addition to or in place of instruction the system itself recommends. A schematized version of the interface is presented in Figure 1. The state of the aircraft system, including the fault to be isolated and any changes brought about by user actions, is modeled by HYDRIVE's *system model*. In a manner described in greater detail below, the student's performance is monitored by evaluating how he or she uses available information about the system to direct troubleshooting actions. HYDRIVE's *student model* is used to diagnose the quality of specific troubleshooting actions, and to characterize student understanding in terms of more general constructs such as knowledge of systems, strategies, and procedures that are associated with troubleshooting proficiency. The general structure of HYDRIVE is presented in Figure 2, with the modules responsible for student modeling highlighted.

[Insert Figures 1 & 2 about here]

Cognitive grounding

An implicit model of student performance must emerge from an understanding of the nature of performance by individuals with different levels of expertise. The rationale for HYDRIVE's design was established through the application of the PARI cognitive task analysis methodology developed in the Basic Job Skills Program of the Armstrong Laboratories (Means & Gott, 1988; Gitomer et al., 1992). These analyses were intended to reveal critical cognitive attributes that differentiate proficient from less-proficient performers in the domain of troubleshooting aircraft hydraulic systems. PARI analysis is a structured protocol analysis scheme in which maintenance personnel are asked to solve a problem mentally, detailing the reasons for their action (**P**recursor), and the **A**ction that they would take. The technician is presented a hypothetical **R**esult and then asked to make an **I**nterpretation of the result in terms of how it modifies understanding of the problem. They are also asked to represent their understanding of the specific aircraft system they are troubleshooting by drawing a block diagram of the suspect system.

Proficiency differences were apparent in three fundamental and interdependent areas, all of which seem necessary for an effective mental model of a system: system understanding, strategic understanding, and procedural understanding.

System understanding. System understanding consists of how-it-works knowledge about the components of the system, knowledge of component inputs and outputs, and

understanding of system topology, all at a level of detail necessary to accomplish necessary tasks (Kieras, 1988). Novices did not evidence appropriate mental models, as represented by the block diagrams they were asked to draw, of any hydraulic system sufficient to direct troubleshooting behavior. The few "models" they did generate usually showed a small number of unconnected components too vague to be of use in troubleshooting (e.g., Figure 3). Experts' models generally evidenced a fuller understanding of how individual components operated within any given system (e.g., Figure 4), even though they did not understand the internal workings of these same components, which they had only to replace.

Experts also demonstrated a principled sense of hydraulic system functioning independent of the specific F-15 aircraft. They seemed to understand classes of components beyond the specific instances found in a particular aircraft or aircraft system, and organized their knowledge hierarchically according to the functional boundaries of the system. At an even higher level, experts also understood the shared and discrete characteristics of flight control and other hydraulic-related aircraft systems. The most important consequence of this type of understanding is that, in the absence of a completely pre-specified mental model of a system, experts are able to construct a mental model using schematic diagrams. They can flesh out the particulars given their basic functional understanding of how hydraulic systems work in the context of the aircraft.

Strategic understanding. Novices did not employ effective troubleshooting strategies. That is, they demonstrated little ability for using system understanding to perform tasks that would allow them to draw inferences about the problem from the behavior of the system (Kieras, 1988). In many cases, the only strategy available to these individuals was to follow designated procedures in technical materials (Fault Isolation Guides, or FIs), even when it wasn't clear that the symptom matched the conditions described therein. While FIs can be useful tools, novices often fail to understand what information about the system a particular FI procedure provides or how it serves to constrain the problem space. Even in those cases where the technician evidences some system understanding, a *serial elimination* strategy, where components adjacent to each other are operated on in order, is frequently used. This strategy allows the technician to make claims only about a single component at a time. A *space-splitting* strategy, in

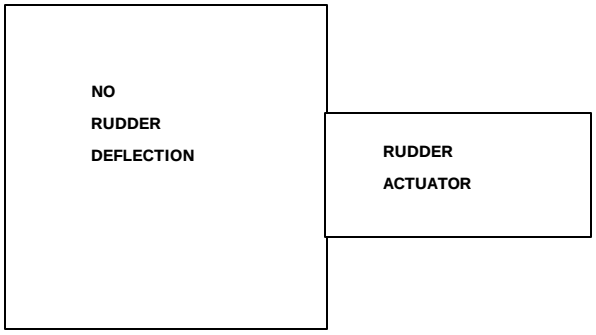


Figure 3
A novice's representation of a flight control problem
produced during the PARI task analysis

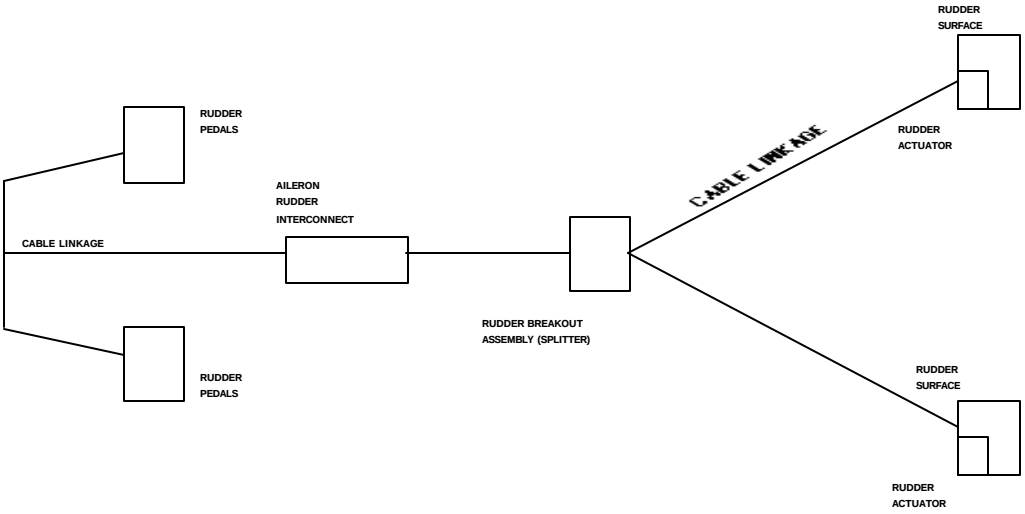


Figure 4
An expert's representation of a flight control problem
produced during the PARI task analysis

contrast, dictates the use of actions that provide information about many components at one time, making this type of strategy much less costly.

Experts try to use space-splitting strategies, isolating problems to a subsystem by using relatively few and inexpensive procedures that can rule out large sections of the problem area. They usually attempt to identify and eliminate power system failures first (e.g., functional failure due to a blown fuse), then activate different parts of the system to find the path along which the failure is manifest; and finally localize the failure to a specific segment of this path (i.e., mechanical, electrical, hydraulic). Experts evaluate results in terms of their mental models of the system and make determinations of the integrity of different parts of the aircraft. When experts

consult the FI guide, they do so as a reference to check whether they may be overlooking a particular problem source, and any FI action is immediately interpreted in terms of and integrated with their system mental model.

Technicians with intermediate skills are quite variable in their use of strategies. When individuals have fairly good system understanding, they frequently evidence effective troubleshooting strategies. When system understanding is weak though, technicians often default to FI and serial elimination strategies. When an intermediate demonstrates a basic understanding of troubleshooting strategy that is dependent on system understanding, the implication for instruction is to focus on system understanding. The evidence suggests that direct strategy instruction may also be necessary for novices.

Procedural understanding. Every component can be acted upon through a variety of procedures which provide information about some subset of the aircraft. Information about some types of components can only be gained by removing and replacing (R&R) them. Others can be acted upon by inspecting inputs and outputs (electrical, mechanical, and/or hydraulic), and by changing states (e.g., switches on or off, increasing mechanical input, charging an accumulator). Some actions provide information only about the component being acted upon, while other actions can provide information about larger pieces of the problem area under certain states of the system model. R&R procedures tend to provide information only about the component being operated upon.

As individuals gain expertise, they develop a repertoire of procedures that can be applied during troubleshooting. Novices are generally limited to R&R actions and the procedures specified in the FI. They often fail to spontaneously use the information that can be provided from studying gauges and indicators and conventional test equipment procedures. Experts are particularly adept at partially disabling aircraft systems and isolating major portions of the problem area as functional or problematic. For example, rudders can be controlled through electrical and/or mechanical inputs. Disabling the electrical system can provide valuable information about the hydraulic and mechanical paths.

The relationship between system, strategic, and procedural understanding. A mental model includes information not only about the inputs and outputs of components, but also available actions that can be performed on components. The tendency to engage in certain procedures or strategies is often a function of the structure and completeness of system understanding, rather than the understanding of strategies or procedures in the abstract. A

student's failure to execute a space-splitting action may appear at first blush to be a strategic failure, but the difficulty may lie with an impoverished understanding of the subsystem—a distinct possibility if the student has exhibited strong strategic practice on other problems for which good system understanding exists.

This view of troubleshooting expertise has implications for instruction as well as for inference. HYDRIVE's instruction focuses on effective system understanding and troubleshooting strategies rather than on optimizing actions to take at a given point in a problem. Ineffective actions raise doubts about a student's system understanding, which might suggest instruction targeted towards student construction of appropriate and useful system models. A key instructional strategy is to help students develop a hierarchical model of system understanding that is the critical feature of expert knowledge. HYDRIVE attempts to make this structure explicit through the use of hierarchical diagrams and similarly-organized verbal information.

Considerations for student modeling

Wenger (1987) describes three levels of information that student modeling might address. Early ITSs focusing on the *behavioral level* were usually concerned with the correctness of student behaviors referenced against some model of expert performance. For example, SOPHIE-I (Brown, Burton & Bell, 1975) contrasted student behaviors with domain performance simulations as a basis for offering corrective feedback. The *epistemic level* of information is concerned with particular knowledge states of individuals. The SHERLOCK ITS (Lesgold, Eggan, Katz, & Rao, 1992) makes inferences about the goals and plans students are using to guide their actions during problem solving. Feedback is meant to respond to "what the student is thinking." The *individual level* addresses broader assertions about the individual that transcend particular problem states. Whereas the epistemic level of diagnosis might lead to the inference that "the student has a faulty plan for procedure X", the individual level of information might include the assertion that "the student is poor at planning in contexts A and B."

This individual level of information has received the least attention in the ITS field. In contrast, educational testing has focused mainly on the individual level, with little explicit attention to the epistemic level. It might be asserted, for example, that an individual has "high ability in mathematics," without an account of the epistemic conditions that characterize high ability. By bridging between the individual and epistemic levels of information, a student model

can have both the specificity to facilitate immediate feedback in a problem-solving situation, and the generality to help sequence problems, moderate instruction, and track proficiency in broad terms. HYDRIVE aims to support generalized claims about aspects of student troubleshooting proficiency with detailed epistemic analysis of particular actions within the system.

The foregoing sections addressed a structural aspect of the inferential task, outlining a rationale for the nature and the grainsize of a student model for a hydraulics-troubleshooting ITS. For any given student working through a HYDRIVE problem, the assessment task is to reason from the student's actions to implications in the student-model space. This problem is harder than the analogous one faced in traditional educational assessment, since there one can devise a collection of predetermined observational settings with predetermined response categories that can be presented to any or all students (i.e., test items). Constraining observations in this manner limits what can be learned, but it is easy to know how to 'score' student's responses. In a relatively unconstrained ITS such as HYDRIVE, however, students can take an unlimited number of routes through a problem; there are no clearly defined and replicable 'items' to score and calibrate. Different students carry out different sequences of action under different system-model configurations; each action depends on multiple aspects of competence, intertwined throughout the diverse situations students leads themselves through. We must, in some fashion, attempt to capture key aspects of their performance in terms of the theory of performance that emerged from the cognitive analysis.

Probability-Based Inference

> *In any inference task our evidence is always incomplete, rarely conclusive, and often imprecise or vague; it comes from sources having any gradation of credibility. As a result conclusions reached from evidence having these attributes can only be probabilistic in nature.*
>
> Schum, 1994, p. xiii.

Inference is reasoning from what we know and what we observe to explanations, conclusions, or predictions. We always reason in the presence of uncertainty. The information we work with is typically incomplete, inconclusive, and amenable to more than one explanation (Schum, 1994). We attempt to establish the weight and coverage of evidence in what we observe, as they inform the inferences and decisions we wish to make. Workers in every field have had to address these questions as they arise with the kinds of inferences and the kinds of evidence they customarily address. Currently, the promise of computerized expert systems has sparked interest in principles of inference at a level that might transcend the particulars of fields

and problems. Historically, this quest has received most attention in the fields of statistics (unsurprisingly), philosophy, and jurisprudence. We shall focus on the concepts and the uses of probability-based reasoning—in particular, mathematical or Pascalian (after Blaise Pascal) probability, from what is usually called a subjectivist (de Finetti, 1974) or personalist (Savage, 1961) perspective.

It is well known that Pascal's trailblazing application of the tools of mathematics to reasoning under uncertainty was sparked by a friend's request for advice on games of chance. Pascal, followed by Bernoulli, Laplace, and others, laid out a framework for reasoning in such contexts. A "random variable" X is defined in terms of a collection of possible outcomes (the sample space), and a mapping from events (subsets of the sample space) to numbers which correspond to how likely they are to occur (probabilities). Probabilities satisfy the following requirements: (i) an event's probability is greater than or equal to 0, (ii) the probability of the event that includes all possible outcomes is 1, and (iii) the probability of an event defined as the union of two disjoint events is the sum of their individual probabilities (Kolmogorov, 1950). We will denote by $p(x)$ the mapping from a particular value x of X onto a probability. These simple axioms lead to consistent inference even for very complex situations, such as games with unknown probabilities linked in complicated ways or with events whose probabilities depend on the outcomes of earlier observations (a form of "conditional" probabilities, or the probability of x given that another variable Z takes the value z, denoted $p(x|z)$)—all of which can be verified empirically because the games can actually be played many times and the frequencies of various events tabulated.

The propriety of Pascalian probability for these aleatory, or chance, situations, is unquestioned. However, as Schum (1994) notes, "there has been lingering controversy, often quite heated, about the extent to which we should accept the Pascalian system in general and Bayes's rule in particular as guides to life in probabilistic inference, especially when our evidence and hypotheses refer to singular or unique events whose probability can rest on no overt enumerative process. … On the epistemic view a probability simply grades the intensity of a person's belief about the likeliness of some event based upon whatever evidence this person has to justify this belief. The issue is: Should all epistemic gradations of probabilistic belief conform to Pascalian rules?" (p. 222). The personalistic Bayesian responds affirmatively without hesitation (we tender our own reservations in the Discussion); when one represents his or her beliefs about a real-world situation in the form of probability distributions, the rules of Pascalian probability ensure that these individual beliefs are consistent with one another, or "coherent." This is particularly important when one contemplates revising beliefs in response to new

information. These constraints permit coherent reasoning from evidence about any subset of variables, to implications for any others, and explicate the manner by which particular pieces of evidence influence beliefs.

Though out of fashion throughout much of this century—a period of great advances in the formalism and methodology of probabilistic reasoning—the view of probability as a personal degree of belief co-existed with its aleatory interpretation since its very beginnings. Leibniz, for example, was concerned with the relational character of probability judgments, from evidence observed to proposition inferred. "This is a natural consequence of his starting point, namely the law," notes Hacking (1975, p. 135); "All legal judgment is *ex datis*." The real question is not whether probability-based reasoning is permissible in applications that lie outside the realm of repeatable chance situations, but whether it is useful; specifically, whether the extent to which the salient aspects and relationships in a given real-world problem can be approximated in this framework, and, if so, whether the required calculations are tractable. The following sections address issues encountered in defining variables, expressing their interrelationships, constructing conformable probability distributions, and carrying out inference, illustrated in the context of HYDRIVE. More recently developed methods of reasoning under uncertainty are mentioned in connection with constraints or difficulties one encounters when applying probability in these contexts.

Kinds of Inference

It is useful to distinguish three kinds of reasoning, all of which play essential and interlocking roles in an ITS (Schum, 1987). *Deductive reasoning* flows from generals to particulars, within an established framework of relationships among variables; i.e., from causes to effects, from diseases to symptoms, from a student's knowledge and skills to observable behavior. *Inductive reasoning* flows in the opposite direction, also within an established framework of relationships—from effects to possible causes, from symptoms to probable diseases, from students' solutions or patterns of solutions to likely configurations of knowledge and skill. Given outcomes, what state of affairs may have produced them? *Abductive reasoning* proceeds from observations to new hypotheses, new variables, or new relationships among variables. The strategy employed in Bayesian inference networks is to erect a reasoning structure in terms of deductive relationships, which, if Pascalian requirements are satisfied, supports conformable inductive inference (Pearl, 1988; Shachter & Heckerman, 1987).

The theories and explanations of a field suggest the structure through which deductive reasoning flows—the "generative principles of the domain," as Greeno (1989) phrased it. The structure for the HYDRIVE system model is deterministic, emanating from the hydraulic, mechanical, and electrical interconnections among the aircraft components. What will the rudder do if I move the control stick when the shear pin is broken? The structure for the student model is probabilistic, emanating from the cognitive analyses described above. For example, given a student who is fairly familiar with troubleshooting strategies and the hydraulics system, but hazy about the workings of the landing gear system, what are the chances of the various possible actions for a given state of a canopy failure? Inductive reasoning flows through these same structures, but in the other direction; the problem is to speculate on circumstances which, when their consequences are projected deductively, lead plausibly to the evidence at hand.

Our modeling objective is to define variables and interrelationships that approximate this structure. We next describe two representative variables suggested by the cognitive analyses, define a relationship between them in terms of conditional probabilities, and use this simple example to illustrate probability-based deductive and inductive inference. Bayes theorem and the concepts of conditional dependence and independence are introduced in this connection. This will be followed by a discussion of how more complex interrelationships among many variables are represented in Bayesian inference networks.

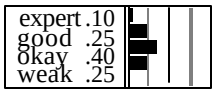
Defining Variables In HYDRIVE

Unlike bridge hands and coin flips, most real-world problems do not present themselves us in terms of natural, ready-made "random variables." It is useful to dispense at the outset the notion that random variables are features of the world; they are rather, features of our representations of the patterns in terms of which we organize our thinking about the world (Shafer, 1988). The subjectivity that non-Bayesian statisticians protest in the specification of personal probabilities is dwarfed by the subjectivity in mapping practically any real-world situation into any formal reasoning framework. From unique events, we must create abstractions which capture aspects we believe are salient but neglect infinitely many others. We must choose the level of detail at which variables will be defined, relationships will be modeled, and analyses will be carried out (Schum, 1994, p. 5; Kadane & Schum, 1992). Although texts on probability or statistics *start* with predefined random variables, the step of conceptualizing our problem in terms of variables amenable to probabilistic inference (particularly "observable variables") was one of the toughest challenges we faced!

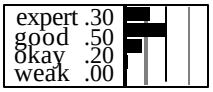
"Strategic knowledge," for example, is a clear abstraction—a shorthand that instructors use to summarize patterns of trainees' behavior, not just troubleshooting actions, but in their conversations, classroom activities, and interactions with instruction (see Pearl, 1988, p. 44, on the very human drive to invent such constructs to organize and explain our experience). Simply stated, as trainees gain competence they tend to take space-splitting actions increasingly often, and when they don't, use serial elimination informatively; they take fewer redundant or irrelevant actions. We might therefore propose a variable called "strategic knowledge" for our student model, with possible values that represent increasing levels of expertise. To illustrate graphic conventions we will be using, Figure 5 depicts three possible states of belief we might have about a student's "strategic knowledge." The first panel represents our belief about a new student entering our course, reflecting our experience that most entering students tend to be relatively weak in troubleshooting strategies. The second panel represents strong belief that a student is fairly good at troubleshooting strategies, perhaps as a result of having studied his transcript, read his supervisor's recommendation, or observed a series of expert-level troubleshooting actions in HYDRIVE. The third panel represents certainty that the student's level of expertise is "weak." Although a state of knowledge is never known with certainty, such an assumption proves useful to reason from in a "what if?" manner when structuring our knowledge about a domain. Later, we will pin down the meaning of "strategic knowledge" by specifying the tendencies of actions we might expect in various troubleshooting situations from a student at each level, moderated by other student-model variables such as subsystem and procedural knowledge.[2] These specifications will represent deductive reasoning, from individual-level variables in the student model to interpretations of actions.

---

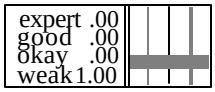
[2] Note that "strategic knowledge" defined in this manner more closely resembles Zadeh's (1983) "fuzzy variables" than the stereotypical notion of variables in probability-based reasoning, as captured by the words of the statistician M.S. Bartlett: "By statistical data and statistical phenomena I refer to the numerical and quantitative facts about groups or classes of individuals or events, rather than facts about the individuals themselves" (quoted in Savage, 1977, p. 5). It is not the nature of variables that determines whether uncertain reasoning is probability-based, but whether the rules of probability are followed in expressing and manipulating beliefs about them.



StratgcKnow



StratgcKnow



StratgcKnow

Figure 5
Three configurations representing possible belief about "Strategic Knowledge"

As a second variable, we define such a lower-level abstraction, an “interpreted action” in a given situation in a problem being worked. As potential observable variables, these action episodes are not predetermined and uniquely-defined in the manner of usual assessment items, since a student could follow a virtually infinite number of paths through the problem. Rather than attempting to model all possible system states and specific possible actions within them, HYDRIVE posits equivalence classes of states, each of which could arise many times or not at all as a given student works through a problem. Let us first consider the definition of a prototypical “interpreted action,” and second, its relationship to a single prototypical “strategic knowledge.”

“Interpreted actions” constitute the interface between the individual-level variables of HYDRIVE’s student model and the virtually unique sequences of actions that individual students take as they work through a problem. They are the lowest level of probability-based reasoning in HYDRIVE. The input to these variables corresponds to ‘observed data’ for probabilistic reasoning, although they are actually fallible judgments from a rule-based parsing of students’ actions. This is referred to as “virtual evidence” in the expert systems literature (e.g., Neapolitan, 1990, p. 230), and highlights a source of uncertainty that we will have to take into account. The demarcation between the probability-based reasoning and the rule-based

reasoning we will now describe is analogous to that between the subconscious feature-detection and pattern matching of human perception, and the comparatively ponderous and labored conscious analysis we then apply—comparing, rationalizing, finding explanations. "Both conscious and unconscious modes of thought are powerful and essential aspects of human life. Both can provide insightful leaps and creative moments. And both are subject to errors, misconceptions, and failures" (Norman, 1988, pp. 125-126).

The values of "interpreted action" variables are produced by HYDRIVE's *system model*, *action evaluator*, and *strategy interpreter* . A student's actions are evaluated in terms of the information they yield in light of the current state of the system model. The action evaluator calculates the effects on the problem area of an action sequence the student performs. The strategy interpreter makes rule-based inferences about the student's apparent strategy usage based on the nature and the span of problem area reduction obtained from the action evaluator.

The *system model* appears to the student as an explorable, testable aircraft system in which a failure has occurred. It is built around sets of components connected by inputs and outputs. Connections, or "edges," are expressed as pairs of components, the first producing an output which the second receives as an input, qualified by the type of power characterizing the connection. The connection between a rudder and its actuator (the servomechanism which causes it to move) would be "left rudder servocylinder_left rudder (mechanical)" because the actuator produces a mechanical output which the rudder processes as input. The output of a component is determined by its inputs and the internal state of the component. A failure may cause no output or an incorrect output to be produced. Every component also has actions that can be performed on it. Some can be set or manipulated (e.g., switches or control handles), others can be checked for electrical function (e.g., relays), and others can be inspected visually (e.g., mechanical linkages). The system model processes the actions of the student and propagates sets of inputs and outputs throughout the system. A student activates the system model by providing input to the appropriate components, and can then examine the results for any other component of the system. A student can move the landing gear handle down and then observe the operation of the landing gear. If the landing gear does not descend, the student may decide to observe the operation of other components to begin to isolate the failure.

The *action evaluator* considers every troubleshooting action from the student's point of view, in terms of the information it conveys about the problem area. All components in the system are part of the initial problem area, represented as sets of input/output edges. When a

student acts to supply power and input to the aircraft system, the effects of this input spread throughout the system model (as values propagated along a continuum of component edges), creating explicit states in a subset of components—the active path, comprised of the points from which input is required to initiate system function to its functionally terminal outputs, and all the connections in between. The action evaluator updates its problem area as if the student correctly judged whether observations reveal normal or abnormal component states. If, having supplied a set of inputs, a student observes the output of a certain component that the system model 'knows' is normal, then it is possible for the student to infer that all edges on the active path, up to and including the output edge, are functioning correctly and remove them from the problem area. If the student in fact makes the correct interpretation and draws the appropriate inferences, then the problem areas that the student model and the student hold will in fact correspond and troubleshooting continues smoothly. But if the student decides that the observed component output was unexpected, or abnormal, then, at least in the student's mind, all the edges in the active path would remain in the problem area, any others would be eliminated, and the problem area maintained by the student model would begin to diverge significantly from the one present in the student's mind; irrelevant and redundant actions become more likely.

The *strategy interpreter* evaluates changes to the problem area (denoted as **k**), or the entire series of edges belonging to the system/subsystem where the problem occurs. As a student acts on the system model, **k** is reduced because the results of action sequences, if correctly interpreted, eliminate elements as potential causes of the failure. If the student inspects any particular component, the system model will reveal a state which may or may not be expected from the student's perspective. HYDRIVE employs a relatively small number of strategy interpretation rules (~25) to characterize each troubleshooting action in terms of both the student and the best strategy.[3] An example of a student strategy rule is:

> *IF an active path which includes the failure has **not** been created and the student creates an active path which does **not** include the failure and the*

---

[3] These rules can be generalized to other troubleshooting domains. The generalizability resides in the ability to explicitly define strategies in terms of an action's effect on **k**. While other domains may require the definition of strategies different from the one used by HYDRIVE, as long as these strategies can be referenced to changes in the state of **k**, or some similar representation, such generalization is quite straightforward.

*edges removed from **k** are of one power class, THEN the student strategy is splitting the power path.*

Deductive and Inductive Reasoning

Glenn Shafer points out that "probability is not really about numbers; it is about the structure of reasoning" (quoted in Pearl, 1988, p. 77). The crux of probability-based reasoning is sorting out how what we belief or what we observe influences other things we might believe or might expect to observe. We can use two variables as we have just defined to illustrate the interplay of deductive and inductive reasoning in a framework grounded by the abductive reasoning of the cognitive analyses. For now we will assume that the student in question has strong knowledge of the problematic subsystem and relevant procedures; we will consider later how the influence of these additional factors is brought into play.

Figure 6 shows the flow of deductive reasoning. Consider a scenario near the end of a problem solution, where space-splitting is no longer an option. What are our expectations that a student at each level of strategic knowledge might perform action sequences interpreted as "serial elimination," 'redundant action," "irrelevant action," and "remove and replace"? Serial elimination is the best strategy available; remove and replace is useful but not efficient; both redundant and irrelevant actions are undesirable. Each panel depicts conditional probabilities of the various action categories, given level of strategic knowledge (Table 1 gives numerical values for this illustration). We see increasing likelihood for serial elimination and decreasing likelihood of redundant and irrelevant actions as level of knowledge increases—although even expert sometimes make redundant moves, and novices make what are interpretable expert moves, if not always for the right reasons.

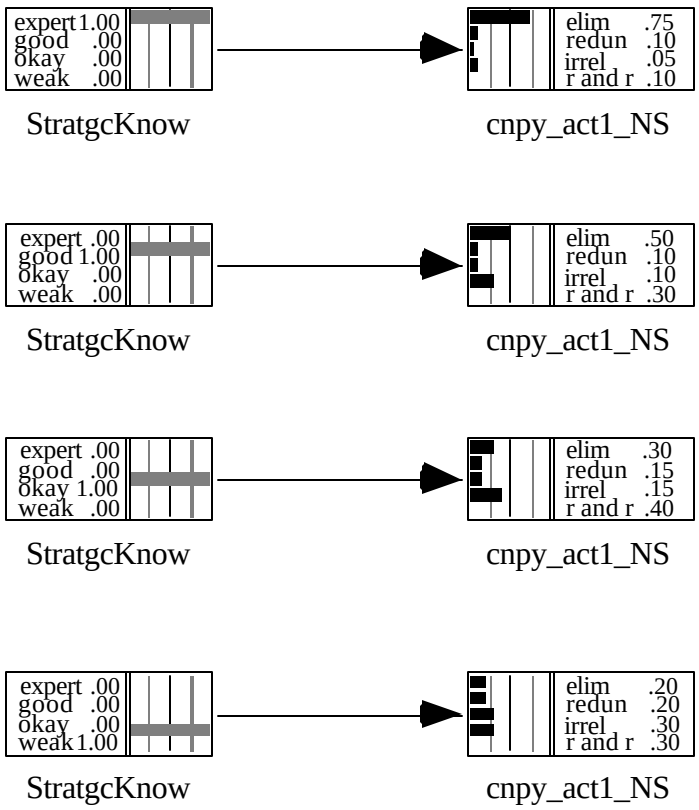


Figure 6
Conditional probablilities of interpreted action sequences, given Strategic Knowledge

Where do these probabilities come from? Initial values were set based on qualitative input from expert instructors, patterns observed in PARI traces, and modifications based on "reasonableness checks" from simulated inputs and outputs. Current research in probability-based reasoning concerns modeling the sources of information about these conditional probabilities, and the sensitivity of inferences to errors or misspecification in them (Schum, 1994, p. 188). Moreover, the probability framework allows conditional probabilities to be characterized as unknown parameters—another level of modeling to represent our beliefs about the structures of relationships among observable and student-model variables—which can capture the "vagueness" of our beliefs about them, yet be coherently updated and made more precise as experience accumulates (Spiegelhalter & Cowell, 1992). Whereas Table 1 simply provided numerical values for the conditional probabilities in the example, then, a more complete representation of belief would take the form of a probability distribution for these conditional probabilities, which itself depended on other aspects of knowledge and information.

Table I
Numerical Values of Conditional Probabilities of Interpreted Action Sequences, Given Strategic Knowledge

	Conditional Probability of Interpreted Action Sequence			
Strategic Knowledge	Serial Elimination	Redundant Action	Irrelevant Action	Remove and Replace
Expert	.75	.10	.05	.10
Good	.50	.10	.10	.30
Okay	.30	.15	.15	.40
Weak	.20	.20	.30	.30

In practice, we must reason inductively; in this case, from interpreted actions to updated beliefs about the student's strategic knowledge. This is accomplished in probability-based reasoning by means of Bayes theorem. Let x be a variable whose probability distribution $p(x|z)$ depends on the variable z. Suppose also that prior to observing x, belief about the value of z can be expressed in terms of a probability distribution $p(z)$ For example, we may consider all possible values of z equally likely, or we may have an empirical distribution based on values observed in the past. Bayes Theorem says

$$p(z|x) = \frac{p(x|z)p(z)}{p(x)}, \tag{1}$$

where $p(x)$ is the expected value of x over all possible values of z—a normalizing constant required by the Pascalian axiom that our belief about z after having learned x must also be represented by a probability distribution that sums to one.[4] To illustrate, suppose we start from the initial new-student beliefs about strategic knowledge depicted in the first panel in Figure 7, and observe one action in the scenario with the (deductive) expectations depicted in Figure 6. If we observe an action interpreted as serial elimination (again recalling that this is not necessarily what the student had in mind) and apply Bayes theorem, we obtain the results in the first panel of Figure 7. We maintain the direction of the arrow because this was the direction in

[4] Dempster-Shafer belief theory (Shafer, 1976) extends Bayesian inference in a manner that can also withhold support from all or some possibilities without having to assign support to other possibilities.

which we specified conditional probabilities. Similar calculations would lead to the results in the other panels if we had observed any of the other possible interpretations.

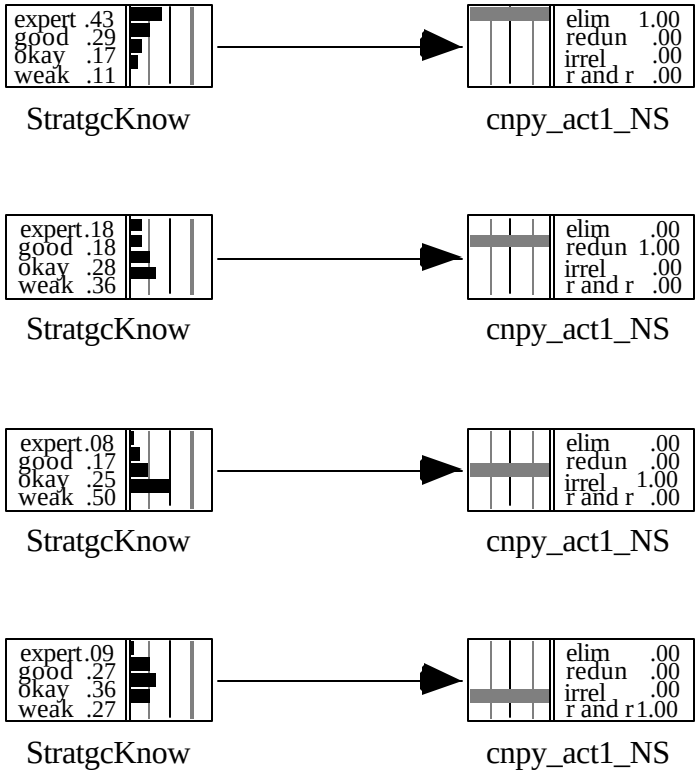


Figure 7
Updated probablilities for Strategic Knowledge, given interpreted action sequences

This last point deserves emphasis, for it is the essence the characterization of belief and weight of evidence under the paradigm of mathematical probability (Good, 1950):

- Before observing a datum x, belief about possible values of a variable Z is expressed as a probability distribution, namely, the prior distribution $p(z)$. (This "prior" distribution can be conditional to other previous observations, and belief about Z may have been revised many times previously; the focus here is just on change in belief associated with observing x, *ceteris paribus*. This capacity to incorporate new information as it arrives suits the ITS context.)

- Posterior to observing the datum x, belief about possible values of Z is expressed as another probability distribution, the posterior distribution $p(z|x)$.

- The evidential value of the observation x is conveyed by the multiplicative factor that revises the prior to the posterior for all possible values of Z, namely, the likelihood function $p(x|z)$. One can examine the *direction* by which beliefs associated with any given z change in response to observing x (is a particular value of z now considered more probable or less probable than before?) and the *extent* to which they change (by a little or by a lot?).

Bayesian Inference Networks

> *What is wanted is simple enough in purpose,—namely, some method which will enable us to lift into consciousness and state in words the reasons why a total mass of evidence does or should persuade us to a given conclusion, and why our conclusion would or should have been different or identical if some part of that total mass of evidence had been different.*
>
> Wigmore, 1937, p. 8.

HYDRIVE moves from the space of unique observations to a space of random variables by interpreting action sequences in terms of equivalence classes—a great simplification, to be sure, but further challenges lie ahead. The problem is to evaluate in terms of the student-model variables the meaning of many such actions, some in equivalent scenarios and others not, each involving different subsytems and aspects of strategic understanding, each allowing for the possibility that the interpreter's evaluation does not match the student's thinking. This is analogous to the problem that jurists routinely confront as they must draw inferences from often large volumes of disparate kinds of evidence. In the first third of this century, Dean of Evidence at Northwestern University John Henry Wigmore sought to explicate principles upon which evidence-based inference appeared to be founded in the law. Although every case is unique, he identified recurring patterns in relationships among propositions to be proved and propositions that tend to support or refute them. "Basic concepts include conjunction; compound propositions; corroboration; convergence; and catenate inferences (inference upon inference) … Each of these notions raises difficult questions about what is involved in determining the overall probative force or weight of evidence" (Twining, 1985, p. 182).

Although Wigmore developed a system for charting the structure of arguments to aid understanding of these relationships in particular cases, he did not claim to prescribe rules for determining that outcome; that is, how to combine a mass of evidence into summary judgments, or to characterize its weight. He left it to the jurors to determine, in a Baconian sense, the extent to which a mass of evidence persuades them of the story of the case. Mathematical probability

does provide tools for combining evidence within a substantively-determined structure—provided that the crucial elements of the situation can be satisfactorily mapped into the probability framework. A first requirement is to express the things we wish to talk about in terms of variables, as discussed above in the context of HYDRIVE. A second is to express the substantive, theoretical, or empirical relationships we perceive among them in terms of structural relationships among probability distributions; that is, the basic evidential relationships Wigmore described.

The notions of conditional independence and dependence are critical in this regard. We begin by saying what we mean by independence. Two random variables X and Z are *independent* if their joint probability distribution $p(x,z)$ is simply the product of their individual distributions, or $p(x,z) = p(x)p(z)$. These variables are unrelated, in the sense that knowing the value of one provides no information about what the value of the other might be. Conversely, conditional *dependence* means that one's belief about the likelihood of values of X depends on what one believes about Z, so that $p(x,z) \neq p(x)p(z)$. The nature of the troubleshooting action we expect from a student (X) depends on our belief about the student's level of strategic knowledge (Z); we denote the conditional distribution as $p(x|z)$. This notion is important in applied work because "the inferential force of one item of evidence may depend in very complex ways upon the *background* provided by other items of evidence we have" (Schum, 1994, p. 208). *Conditional independence* means that what one believes about X might depend on what one believes about Z—but not if one already happens to know the value of another variable Y; that is, $p(x,y|z) = p(x|z)p(y|z)$, or X and Z are conditionally independent given Y. The troubleshooting action we observe in one scenario certainly influences what we expect in the next, but we will posit that it would not if we knew with certainty the values of the student-model variables that were required in both scenarios.

As attention shifts to inductive reasoning, applying Bayes theorem in its textbook form (Equation 1) becomes unwieldy quickly as the number of variables in a problem increases. Efficient probability-based inference in complex networks of interdependent variables is an active topic in statistical research, spurred by applications in such diverse areas as forecasting, pedigree analysis, troubleshooting, and medical diagnosis (e.g., Lauritzen & Spiegelhalter, 1988; Pearl, 1988; for an introduction to Bayes nets in cognitive diagnosis, see Mislevy, 1995, and Martin & VanLehn, 1993). Interest centers on obtaining the distributions of selected variables conditional on observed values of other variables, such as likely characteristics of offspring of selected animals given characteristics of their ancestors, or probabilities of disease states given symptoms and test results.

A *recursive representation* of the joint distribution of a set of random variables $x_1,\ldots,x_n$ takes the form

$$p(x_1,\ldots,x_n) = p(x_n|x_{n-1},\ldots,x_1)p(x_{n-1}|x_{n-2},\ldots,x_1)\cdots p(x_2|x_1)p(x_1)$$

$$= \prod_{j=1}^{n} p(x_j|x_{j-1},\ldots,x_1), \qquad (2)$$

where the term for j=1 is defined as simply p(x_1). A recursive representation can be written for any ordering of the variables, but one that exploits conditional independence relationships is useful because variables drop out of the conditioning lists. A graphical representation of (2), or a directed acyclic graph (DAG), depicts each variable as a node; each variable has an arrow drawn to it from any variables on which it is directly dependent (its "parents"). Conditional independence corresponds to omitting arrows ("edges") from the DAG, thus simplifying the topology of the network.

The conditional independence relationships suggested by substantive theory play a central role in the topology of the network of interrelationships in a system of variables. If the topology is favorable, such calculations can be carried out efficiently through extended application of Bayes theorem even in very large systems, by means of strictly local operations on small subsets of interrelated variables ("cliques") and their intersections. Discussions of construction and local computation in Bayesian inference networks can be found in the statistical and expert-systems literature; see, for example, Lauritzen & Spiegelhalter (1988), Neapolitan (1990), and Pearl (1988). Computer programs that carry out the required computations include HUGIN (Andersen, Jensen, Olesen, & Jensen, 1989) and ERGO (Noetic Systems, 1991).

A Simplified HYDRIVE Bayesian Inference Network

Figure 8 is a DAG expressing the dependence relationships in simplified version of the inference network for the HYDRIVE student model. The direction of the arrows represents the deductive flow of reasoning which we use to construct probability distributions that incorporate the depicted dependence structure. A joint probability distribution for all these variables can be constructed by first assigning a probability distribution to each variable which has no parents (in this example, there is only one: "overall proficiency"); then for each successive variable, assigning a conditional probability distribution to its possible values for each possible combination of the values of its parents. The values expressed in these assignments incorporate

such patterns as conjunctive or disjunctive relationships, incompatibilities, and interactions among diverse influences. Four groups of variables can be distinguished: (1) The rightmost nodes are the "interpreted actions," the results of rule-driven epistemic analyses of students' actions in a given situation. Two prototypical sets appear, each corresponding to an equivalence class of potential observables in a give scenario; three members of the class are represented in both cases.[5] (2) The immediate parents of the interpreted action variables are the knowledge and strategy requirements that in each case define the class. (3) The long column of variables in the middle concerns aspects of subsystem and strategic knowledge, which correspond to instructional options. (4) To the left are summary characterizations of more generally construed proficiencies.

The equivalence classes of actions in this figure concern canopy situations in which space-splitting is not possible, and landing gear situations in which space-splitting is possible. Figure 8 depicts belief after observing, in three separate situations from the canopy/no-split class, one redundant and one irrelevant action (both ineffectual troubleshooting moves) and one remove-and-replace (serviceable but inefficient). Serial elimination would have been the best strategy in this case, and is most likely when the student has strong knowledge of this strategy and all relevant subsystems. Remove-and-replace is more likely when a student possesses some subsystem knowledge but lacks familiarity with serial elimination. Weak subsystem knowledge increases chances of irrelevant and redundant actions.

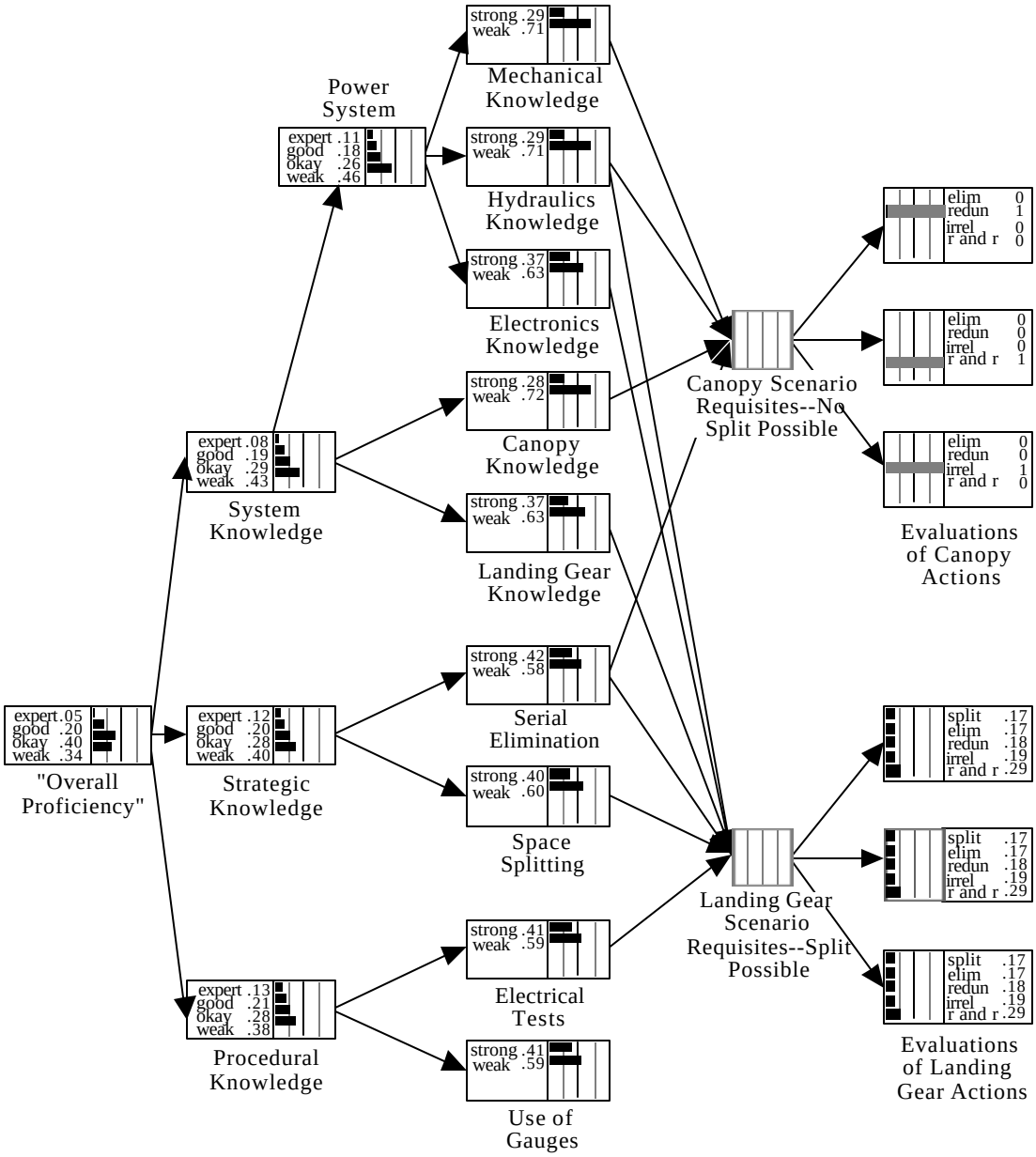
Subsystem and *strategy variables* serve to summarize tendencies in interpreted behaviors at a level addressed by instruction, and to disambiguate patterns of actions in light of the fact that inexpert actions can have several causes. Figure 8 shows the state of belief after observing three inexpert actions concerning the canopy subsystem. Belief is shifted toward lower values for serial elimination, and for all subsystem variables directly involved in the situation (mechanical, hydraulic, and canopy knowledge). Any or all could be a problem, since all are required for high likelihoods for expert actions. Variables for subsystems not directly involved in these situations are also lower, because to varying extents, students familiar with one

---

[5] A given student may, while working through a problem, confront situations from a given equivalence class many times. Mislevy's (1994) algorithm absorbs information from an indefinite number of a class of independent and identically distributed variables, while storing and manipulating only two copies of representative class members.

subsystem tend to be familiar with others, and, to a lesser extent, students familiar with subsystems tend to be familiar with troubleshooting strategies. These relationships are expressed by means of the more *generalized system* and *strategy knowledge* variables at the left of the figure. These variables serve to exploit the indirect information about aspects of knowledge not directly tapped, and to summarize broadly construed aspects of proficiency for evaluation and problem-selection.

Figures 9 and 10 represent the state of belief that would result after observing two different sets of actions in situations involving the landing gear in which space-splitting is possible. Figure 9 shows the results of three more inexpert action sequences. Status on all subsystem and strategy variables is further downgraded, and reflected in the more generalized summary variables. Figure 10 shows the results of observing three good actions: two space-splits and one serial elimination. Belief about strategic skill has increased, as have beliefs about subsystems involved in the landing gear situations. Problems in mechanical and canopy subsystem knowledge are now the most plausible explanations of the three inexpert canopy situation actions. The diffuse belief at the generalized proficiency level results from the uneven profile of subsystem knowledge (a kind of "suspension of belief"; Edwards, 1988), despite fairly accurate information about individual aspects of the student's knowledge.



Note: Bars represent probabilities, summing to one for all the possible values of a variable. A shaded bar extending the full width of a node represents certainty, due to having observed the value of that variable; i.e., a student's actual responses to tasks.

Figure 8
Status of Student Model after Observing Three Inexpert Actions in Canopy Situations

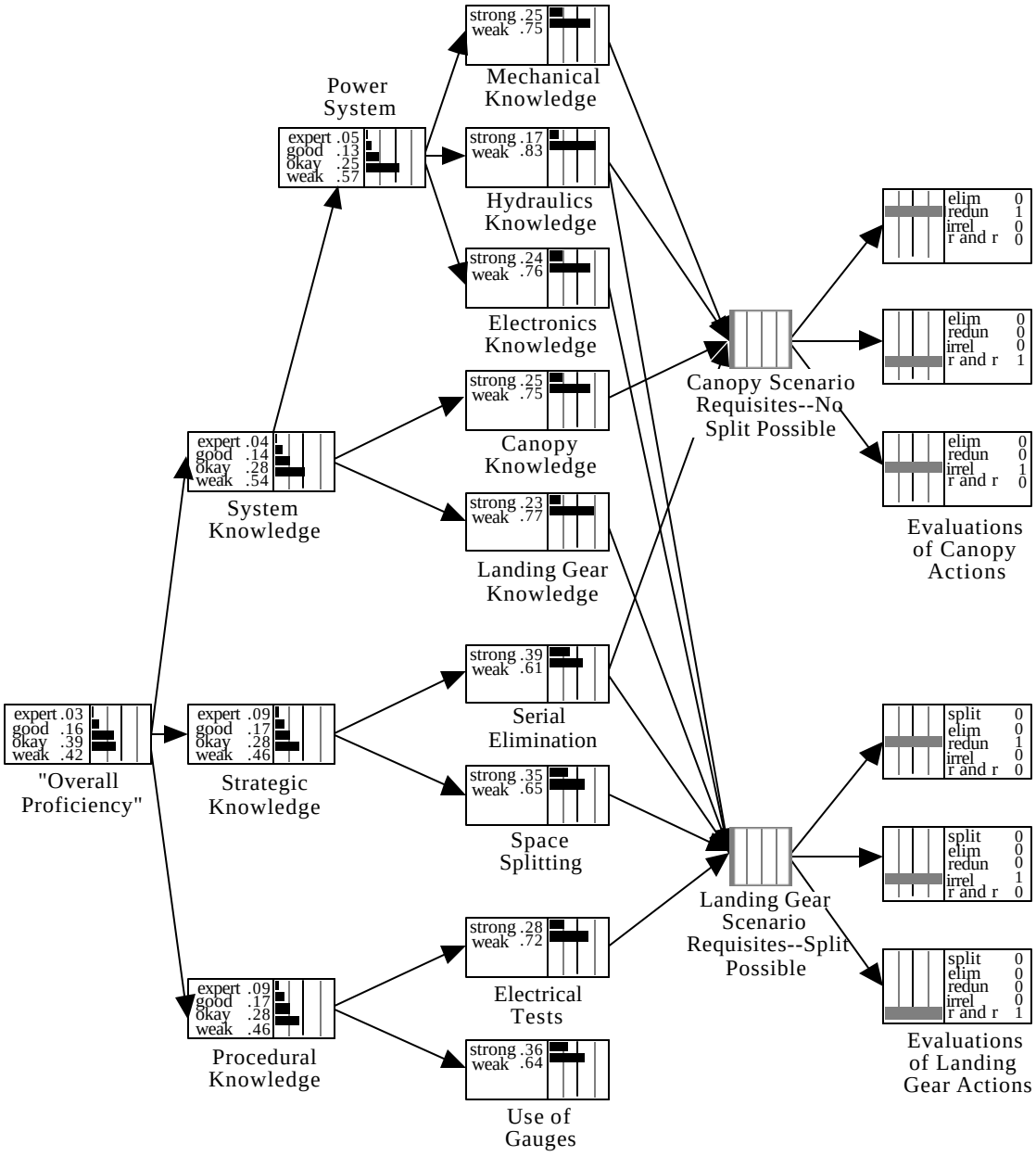
In some problems, it is possible to determine empirically the conditional probabilities of observable variables given student-model variables (e.g. Béland & Mislevy, 1992). In HYDRIVE we do not have the luxury of analyzing large numbers of solutions from

acknowledged experts and novices of various configurations. Initial values were set subjectively, and revised through model-checking activities. The objective is to encode a network structure and conditional probabilities specifications which correspond with experience to date not only locally (i.e., for a single given action-situation) but globally (i.e., after accumulating evidence over a series of actions within a problem, then over a series of problems). Artificial data based on PARI traces were entered to simulate a student working within HYDRIVE, and the behavior of the network was evaluated in light of our expectations from the cognitive analyses. At times, student-model variables were updated too slowly or too rapidly under the initial probabilities as observations accumulated; some updates moved student-model beliefs in unexpected directions. Because all the probabilities are set at the individual node level, the behavior of the entire network is difficult to anticipate. However, by repeatedly applying data, and evaluating the network's behavior, probabilities can be tuned so that the system behaves in a manner consistent with human judgments of performance. This tuning process resembles that required under rule-based and fuzzy-logic systems as well; the only distinction of the probabilistic grounding is that this knowledge engineering is structured so as to assure the satisfaction of the Pascalian axioms. Probabilistic reasoning doesn't dictate what ones beliefs should be about or what they should be; it only demands a certain consistency among those beliefs.

Additional Grounds for Revising Belief

> *[I]ntroducing some model of disease evolution in time, and dealing with treatment, as diagnosis is hard to divorce from therapy in any practical sense.*
>
> Szolovits & Pauker, 1978, p. 128.

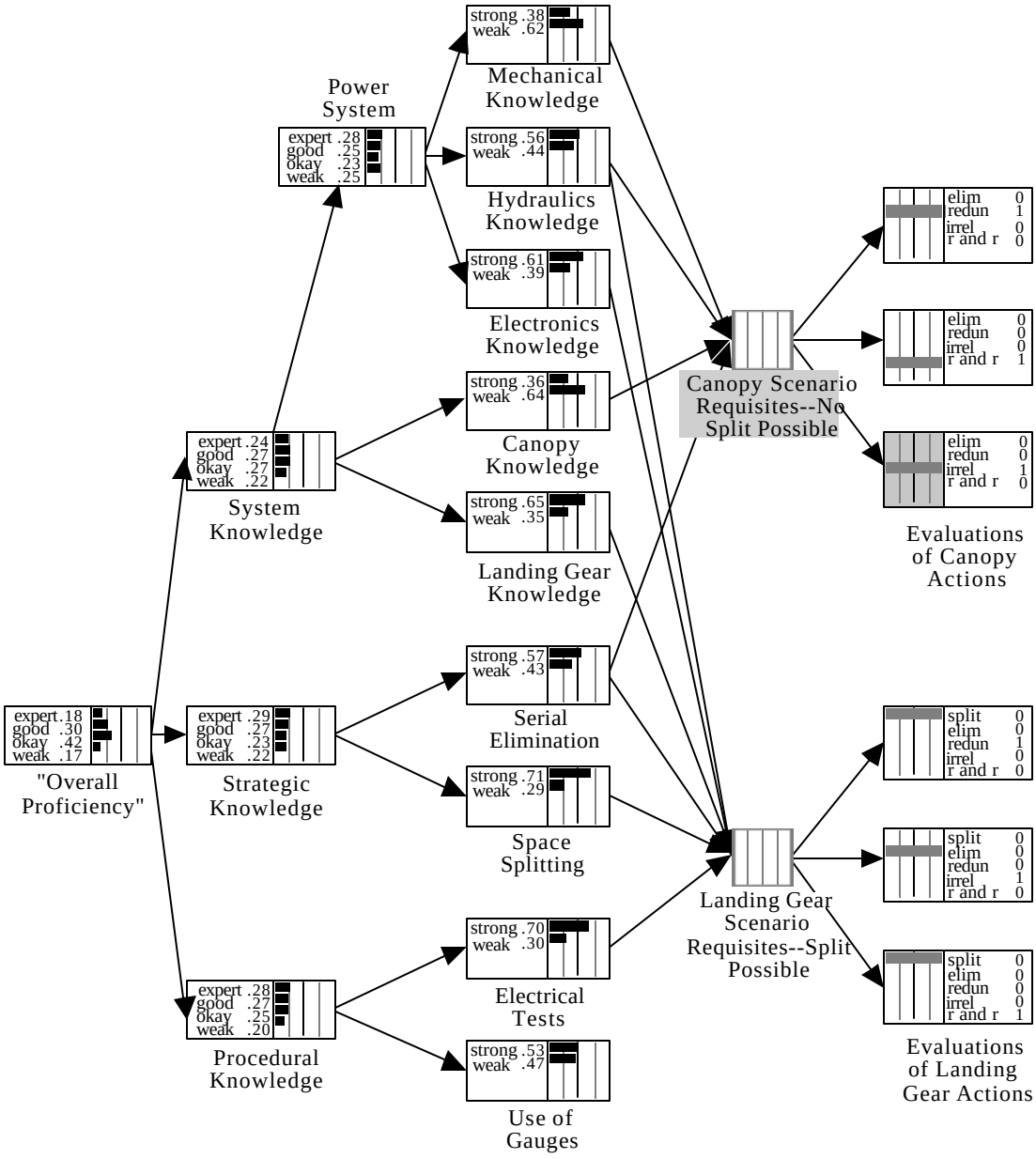
The preceding discussion and examples concerned updating belief about a static student model. That is, even though observations are obtained sequentially over time, they are presumed to simply provide additional information about values of student-model variables that remain constant over time. Most of our work so far has this character, as we have concentrated on modeling proficiencies within self-contained problem exercises. The whole point of an ITS, however, is to help students *change* over time; in particular, to improve their proficiencies. This section concerns two other reasons for modifying belief about student-model variables: changes due to explicit instruction, and changes due to implicit learning. In either case, the requirement under a probabilistic approach is to do so



Note: Bars represent probabilities, summing to one for all the possible values of a variable. A shaded bar extending the full width of a node represents certainty, due to having observed the value of that variable; i.e., a student's actual responses to tasks.

Figure 9

Status of Student Model after Observing Three Inexpert Actions in Canopy Situations and Three Inexpert Actions in Landing Gear Situations



Note: Bars represent probabilities, summing to one for all the possible values of a variable. A shaded bar extending the full width of a node represents certainty, due to having observed the value of that variable; i.e., a student's actual responses to tasks.

Figure 10
Status of Student Model after Observing Three Inexpert Actions in Canopy Situations and Three Expert Actions in Landing Gear Situations

in a manner that maintains coherency. The approach described below accomplishes this end without requiring a full-blown dynamic model to be constructed and maintained.

Updating based on direct instruction. Although HYDRIVE's system model functions as a discovery world for system and procedural understanding from the student's point of view, the evaluations its student modeling components makes are based on an implicit strategic goal structure observed in expert troubleshooting. This structure is made explicit in HYDRIVE's instruction. The student is given great latitude in pursuing the problem solution. Prompts or reminders (i.e., diagnostics) are given only when a student action constitutes an important violation of the rules associated with the strategic goal structure. HYDRIVE recommends direct instruction only when accumulating information across scenarios shifts belief about, say, knowledge of a subsystem or strategy, sufficiently downward to merit more specifically focused feedback, review and exercises. The student is free to follow this recommendation, choose different instruction, or continue troubleshooting without any instruction.

Such directed instruction can be expected to change students' understanding. Whereas updating beliefs about presumably static student-model variables from interpreted actions involved entering findings for the latter and propagating their implications upward through the network, updating belief about changes in student-model variables involves direct manipulation of them, the implications of which are propagated both upward to related aspects of knowledge and downward to revise expectations for future actions. The degree of change is based on the student's performance on the exercises that accompany the instruction. This can be modeled in a small stand-alone Bayesian inference network that embodies a Markovian process for change while incorporating our uncertainty about the exact value of the student-model variable of interest. The probability distribution before instruction and the outcome of the instructional exercises are entered into the network, and modeled beliefs after instruction are output (Figure 11). The conditional probabilities of level of competence after instruction, given level before instruction and performance in exercises (e.g., Table 2), may also be refined over time, starting with expert opinion and limited experience and honed with the results of accumulating experience.

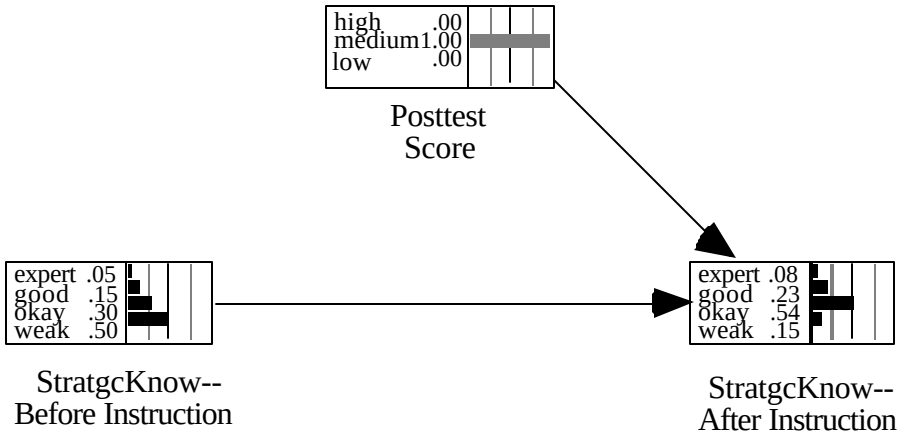


Figure 11

A Markov framework for updatedbelief about Strategic Knowledge due to Direct Instruction

Table II

Conditional Probabilities of Strategic Knowledge after Instruction, Given Probabilities before Instruction and Posttest Performance

Status before Instruction	Posttest Performance	Conditional Probability of Status after Instruction			
		Expert	Good	Okay	Weak
Expert	High	1.00	.00	.00	.00
	Medium	.95	.05	.00	.00
	Low	.90	.10	.00	.00
Good	High	.70	.30	.00	.00
	Medium	.20	.75	.05	.00
	Low	.05	.85	.10	.00
Okay	High	.20	.70	.10	.00
	Medium	.00	.30	.70	.00
	Low	.00	.15	.80	.05
Weak	High	.05	.55	.40	.00
	Medium	.00	.05	.65	.30

Low	.00	.00	.20	.80

Updating based on learning while problem-solving. Even without direct instruction, students can be expected to improve their troubleshooting skills as a result of practicing them and thinking through the problems. Although this probably occurs in increments throughout any given problem, we follow an expedient employed by Kimball (1982) in a calculus tutor: Revisions to belief associated with implicit learning are effectuated only between problem boundaries. Kimball's tutor, like John Anderson's LISP tutor (Anderson & Reiser, 1985), revises belief in a manner consistent with probability axioms through an explicit learning model, a la Estes (1950). That is, a particular functional form for change is presumed, and degree of learning must also be assumed or estimated. We employ a more conservative and less model-bound approach, which might be flippantly called a "forgetting" model as opposed to a "learning" model.

The basic idea is to enter each problem with student-model variable distributions that generally agree with the final values from the previous problem as to direction and central tendency, but are more diffuse and thus easier to change in light of new actions driven by possibly different (presumably improved) values. Mislevy (1994b) outlines two strategies for accomplishing this end: downweighting the influence of actions as they recede in time, and, between problem sessions, mixing final distributions with noninformative distributions and propagating the revised versions through the network as described for instructional revisions. The rationale behind these "decaying-information estimators" is analogous to that of reducing the influence of outliers in robust estimation (Huber, 1981). They are less efficient than full-information estimators if there is no secular change, or if there is change and it is modeled accurately; but they can provide better approximations when trends do exist than either ignoring it or modeling it incorrectly. A robustified version of a simpler model, as opposed to a more complex model, provides serviceable, if not optimal, approximations across a variety of departures from the basic form.

Discussion

Pascalian probability provides powerful machinery for coherent reasoning about complex and subtle interrelationships—to the extent that one can capture within its framework the key aspects of a real-world situation (what is important, how important things are related, and what one sees or knows tells about what one doesn't see or doesn't know). If this can be accomplished, advantages both conceptual and practical accrue. A Bayes net built around the

generating principles of the domain makes interrelationships explicit and public, so one can not only monitor what one believes, but communicate why one believes it. A model can be refined over time in light of new information, as when originally-subjective conditional probability specifications are updated in light of accumulating data. Able to calculate predictive distributions of any subset of variables given values of any others, one can investigate both deductive and inductive implications of a modeled structure, using hypothetical data to check for fidelity to what one believes, or real data for fidelity to what one observes (see the review by Spiegelhalter, Dawid, Lauritzen, & Cowell, 1993, on model-checking tools for complex networks). It may be painstaking and difficult work to carry out the requisite modeling tasks, but recent progress in calculation, model-building, and model-checking has been explosive (again, see Spiegelhalter et al, *op cit*).

The challenge most significant in any application is channeling the scope of vision from an open-ended universe of human experience, to a closed universe of variables and probability distributions (Shafer, 1988). We experienced this constraint in two ways with HYDRIVE. The first, discussed above, is having to interpret observations or beliefs in terms of variables over which probabilities sum to one. Just how to do this is not obvious in HYDRIVE's relatively unconstrained observational setting; a student might take literally hundreds of different actions at any point in a problem. HYDRIVE's conceptual progenitor, the SHERLOCK ITS (Lesgold et al., 1992) also interprets action sequences in terms of inferred plans. But SHERLOCK changes values of student-model variables according to rules an action triggers pertaining to this inductive purpose only. These updating rules are easier to construct than HYDRIVE's conditional probability structures, because the rules triggered by any observation can be specified without regard to rules for others. On the other hand, implications of student model values for future actions are not addressed, and are more difficult to check conceptually or empirically. An interpreted action in SHERLOCK is an "event" in the everyday sense of the word, and even in the sense of nonmathematical or Baconian probability (Cohen, 1977)—but not in the sense of Pascalian probability. To accomplish this in HYDRIVE, we cast interpreted actions as members of exhaustive and mutually exclusive classes, so that the updating that occurs when a space-split *did* occur depends intimately on the fact that an R&R, a serial elimination, or a redundant or irrelevant action *did not* occur.

The second place the constraints of Pascalian probability can pinch is the presumption that all potential states of the real-world situation can be satisfactorily approximated under the model, relative to the purpose at hand. Shafer (1976) calls modeling the possibilities one will explore "defining the frame of discernment." As an example, students' actions in the QUEST

tutor for electrical circuits (White & Frederiksen, 1987) are interpreted in terms of a progression of increasingly sophisticated mental models. It is easy to conceive of a Bayesian framework for deductive, then inductive, reasoning based on expected actions of students in the various stages. But what if a particular student's conception differs from any of the postulated models? The probabilities that result from the use of Bayes Theorem (and all the more when embedded in a complex network) depend on the posited structure. Only possibilities built into the model can end up with positive probabilities! Apparently precise numerical statements of belief prove misleading or downright embarrassing when it is later determined that the true state of affairs could not even be approximated in the analytic model.[6] As Will Rogers once said, "It ain't so much the things you don't know that hurt you; it's the things you know that ain't so."

Two strategies help address this problem in applied settings. One approach is to augment theoretically-expected unobservable states with one or more "catch-all" states which increase in probability when unexpected patterns arise in observable data. The MUNIN expert system for neuromuscular diseases (Andreassen et al., 1987), for example, includes a disease state called "other," a catch-all class that merely characterizes examinees in terms of a flat likelihood for all symptom patterns. When symptoms appear that are unlike any of the distinctive patterns typical under the disease states explicitly built into the model, the posterior probability for this "other" class increases. Another approach is to calculate indices of model misfit (e.g., Levine & Drasgow, 1982). While carrying out inference within a given probabilistic structure, indices are calculated to indicate how usual or unusual the observed data are under that structure: If higher-level parameters took their most likely values in accordance with the

---

[6] The House Select Committee on Assassinations assigned a 95% probability to the proposition that four shots were fired in the John Kennedy assassination, based on a dictabelt recording of sounds believed to have been recorded from a microphone on a police motorcycle in Dealy Plaza at the time of the incident. The sound patterns constituting the evidence, assumed to be echo impulses of shots during the six critical seconds, did in fact provide a much better match to experimentally-produced patterns for four shots than any other number of shots. But rock drummer Steve Barber discovered, faintly recorded on the dictabelt in the same time interval, words known to be spoken by Sheriff Bill Decker more than a minute after the assassination (Posner, 1993)—an observation that obviated any relationship between the putative echo impulses and the actual number of shots. The lesson is that the utility of numerical probabilities calculated within a posited inferential structure depends on the structure's fidelity to the real-world situation in question.

observed datum, how likely would this datum be? Surprising observations are flagged, for in such cases actual circumstances may differ most severely from modeled circumstances. Using either of these approaches, a system can flag patterns of evidence that are not likely under *any* of the possibilities built into the model, effectively crying out for further abductive reasoning.

From this perspective, it may not be necessary or even desirable to attempt to exhaustively build all possible conjectures into one all-encompassing network. Shafer (1976, 1988) points out that in many inferential problems, frames of discernment often evolve over time as we accumulate evidence. We add possibilities, refine others, abandon still others. In conjunction with reasoning structures built around "generative principles of the domain," this constructive perspective represents a rapprochement between mathematical probability-based reasoning and cognitive schema-based reasoning (e.g., Pennington & Hastie, 1991). Frameworks of probability-based reasoning aid our understanding of how available information informs current thinking, without claiming finality or "truth" at any stage; rule-based, inductive, and intuitive reasoning aid our construction and improvement of those frameworks. Encountering a wide variety of disparate reasoning problems in our attempt to build a workable ITS and calling upon a complementary combination of reasoning approaches to deal with them, we arrived empirically at a perspective expressed by Schum (1994):

> *Confronted with alternative views of probabilistic reasoning, a frequently asked question is: Which one is to be preferred? … The different formal systems of probabilistic reasoning each resonate to different attributes of the very rich process of probabilistic reasoning. In my analyses of evidence I have viewed these different formal systems not as normative guides to life but as heuristics for examining evidence from different perspectives or standpoints. (p. 284)*

Conclusion

Widely disparaged in the AI and ITS communities a decade ago, probability-based reasoning has emerged as a viable approach to structuring and managing knowledge in the presence of uncertainty. This is due partly to computational advances such as rapid local updating (Spiegelhalter et al, 1993), but more to conceptual progress—in particular, a confluence of ideas about personal probability (e.g., Savage, 1961; de Finetti, 1974) and the structuring of inference (e.g., Schum, 1994). This progress was spurred by the emergence of alternative frameworks for reasoning in the presence of uncertainty, such as fuzzy sets and fuzzy logic (Zadah, 1965), belief theory (Shafer, 1976), and inductive probability (Cohen, 1977). Whether Pascalian probability *couldn't* be used to deal with the problems these writers

advanced was fiercely contested, but clearly it *wasn't*. It is safe to predict continued rapid progress along statistical lines, which can only increase prospects of its usefulness for intelligent tutoring systems. Also required is progress in two additional lines, namely, cognitive psychology and a certain kind of engineering expertise.

Perhaps the main lesson we take from the HYDRIVE project is the importance of cognitive grounding. Arguing in the abstract about advantages and disadvantages of alternative approaches to managing uncertainty is all well and good, and quite necessary—but in the final analysis, the success of a given application will depend on identifying the key concepts and interrelationships in the domain. Incoherent reasoning with sound substance beats coherent reasoning with inadequate substance, if you must choose between them—but coherent reasoning around sound substance dominates! Two areas of attention are germane in an ITS: (1) understanding about principles of domain and how people learn those principles, so as to structure the student model efficaciously, and (2) what we need to see and how to interpret it in light of students' possible understandings, to as to structure observable variables and their relationship to student-model variables. In HYDRIVE we employed rule-based interpretations to identify critical features from a stream of relatively unstructured observations; Shafer (1987) sees the need for an associative memory mechanism for this purpose, strengthening the analogy to human perception.

The engineering expertise we refer to concerns the interface among the statistics, cognitive psychology, computer science, and instructional science that must come together in a successful ITS. Over time, prototypical approaches for developing ITSs consonant with the principles of these three domains must evolve, in the form of examples, effective approaches to common problems, knowledge elicitation schemes aligned to the anticipated model, and expedients that strike good balances among competing properties such as fidelity and computability. Our experiences with HYDRIVE persuade us that the quest will be arduous, but worthwhile.

Acknowledgements

The examples in this paper are based on the HYDRIVE Intelligent Tutoring System, which has been supported by Armstrong Laboratories of the United States Air Force. We are indebted to Sherrie Gott and her staff for their contribution to this effort. Mr. Mislevy's work was supported in part by Contract No. N00014-91-J-4101, R&T 4421573-01, from the Cognitive Science Program, Cognitive and Neural Sciences Division, Office of Naval Research, and by the National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Educational Research and Development Program, cooperative agreement number R117G10027 and CFDA catalog number 84.117G, as administered by the Office of Educational Research and Improvement, U.S. Department of Education. The views expressed here are those of the authors and do not imply any official endorsement by any organizations funding this work.

References

Andersen, S.K., Jensen, F.V., Olesen, K.G., & Jensen, F. (1989). *HUGIN: A shell for building Bayesian belief universes for expert systems* [computer program]. Aalborg, Denmark: HUGIN Expert Ltd.

Anderson, J.R., & Reiser, B.J. (1985). The LISP tutor. *Byte,* **10,** 159-175.

Anderson, J.R., Boyle, C.F., Corbett, A.T. (1990). Cognitive modelling and intelligent tutoring. *Artificial Intelligence,* **42,** 7-49.

Andreassen, S., Woldbye, M., Falck, B., & Andersen, S.K. (1987). MUNIN: A causal probabilistic network for interpretation of electromyographic findings. *Proceedings of the 10th International Joint Conference on Artificial Intelligence* (pp. 366-372). Milan: Kaufmann.

Béland, A., & Mislevy, R.J. (1992). Probability-based inference in a domain of proportional reasoning tasks. *ETS Research Report 92-15-ONR.* Princeton, NJ: Educational Testing Service.

Brown, J. S., Burton, R. R. & Bell, A. G. (1974). SOPHIE: A sophisticated instructional environment for teaching electronic troubleshooting. *BBN REPORT 2790*. Cambridge, MA: Bolt Beranek and Newman, Inc.

Cheeseman, P. (1985). In defense of probability. In *Proceedings of the* 9^{th} *International Joint Conference on Artificial Intelligence* (pp. 1002-1009).

Cheeseman, P. (1986). Probabilistic versus fuzzy reasoning. In L.N. Kanal & J.F. Lemmer (Eds.), *Uncertainty in artificial intelligence* (pp. 85-102). Amsterdam: North-Holland.

Cohen, L.J. (1977). *The probable and the provable*. Oxford: The Clarendon Press.

de Finetti, B. (1974). *Theory of probability (*Volume 1). London: Wiley.

Edwards, W. (1988). Insensitivity, commitment, belief, and other Bayesian virtues, or, who put the snake in the warlord's bed? In P. Tillers & E.D. Green (eds.), *Probability and inference in the law of evidence* (pp. 271-276). Dordrecht, The Netherlands.

Estes, W.K. (1950). Toward a statistical theory of learning. *Psychological Review*, **57,** 94-107.

Gitomer, D.H., Cohen, W., Freire, L., Kaplan, R., Steinberg, L., & Trenholm, H. (1992). *The software generalizability of HYDRIVE* (Armstrong Laboratories Progress Report). Princeton, NJ: Educational Testing Service.

Gitomer, D.H., Steinberg, L.S., & Mislevy, R.J. (1995). Diagnostic assessment of trouble-shooting skill in an intelligent tutoring system. In P. Nichols, S. Chipman, & R. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 73-101). Hillsdale, NJ: Erlbaum.

Good, I.J. (1950). *Probability and the weighting of evidence.* London: Griffin; New York: Hafner.

Greeno, J.G. (1976). Cognitive objectives of instruction: Theory of knowledge for solving problems and answering questions. In D. Klahr (Ed.), *Cognition and instruction* (pp. 123-159). Hillsdale, NJ: Erlbaum.

Greeno, J.G. (1989). A perspective on thinking. *American Psychologist,* **44,** 134-141.

Hacking, I. (1975). *The emergence of probability*. Cambridge: Cambridge University Press.

Huber, P.J. (1981). *Robust statistics*. New York: Wiley.

Kadane, J.B., & Schum, D.A. (1992). Opinions in dispute: the Sacco-Vanzetti case. In J.M. Bernardo, J.O. Berger, A.P. Dawid, & A.F.M. Smith (Eds.), *Bayesian Statistics 4* (pp. 267-287). Oxford, U.K.: Oxford University Press.

Kieras, D.E. (1988). what mental model should be taught: choosing instructional content for complex engineered systems. In M.J. Psotka, L.D. Massey, & S.A. Mutter (Eds.), *Intelligent tutoring systems: Lessons learned* (pp. 85-111). Hillsdale, NJ: Lawrence Erlbaum.

Kimball, R. (1982). A self-improving tutor for symbolic integration. In D. Sleeman & J.S. Brown (Eds.), *Intelligent tutoring systems*.

Kolmogorov, A.N. (1950). *Foundations of the theory of probability*. New York: Chelsea.

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

Lesgold, A. M., Eggan, G., Katz, S., & Rao, G. (1992). Possibilities for assessment using computer-based apprenticeship environments. In J. W. Regian and V.J. Shute (Eds.), *Cognitive approaches to automated instruction* (pp. 49-80). Hillsdale, NJ: Lawrence Erlbaum.

Levine, M., & Drasgow, F. (1982). Appropriateness measurement: Review, critique, and validating studies. *British Journal of Mathematical and Statistical Psychology,* **35,** 42-56.

Lindley, D.V. (1987). The probability approach to the treatment of uncertainty in artificial intelligence and expert systems. *Statistical Science,* **2,** 25-30.

Martin, J.D., & VanLehn, K. (1993). OLEA: Progress toward a multi-activity, Bayesian student modeler. In S.P. Brna, S. Ohlsson, & H. Pain (Eds.), *Artificial intelligence in*

education: Proceedings of AI-ED 93 (pp. 410-417). Charlottesville, VA: Association for the Advancement of Computing in Education.

Means, B., & Gott, S.P. (1988). Cognitive task analysis as a basis for tutor development: Articulating abstract knowledge representations. In M.J. Psotka, L.D. Massey, & S.A. Mutter (Eds.), *Intelligent tutoring systems: Lessons learned* (pp. 35-58). Hillsdale, NJ: Erlbaum.

Mislevy, R.J. (1994a). Virtual representation of IID observations in Bayesian belief networks. *ETS Research Memorandum 94-13-ONR.* Princeton, NJ: Educational Testing Service.

Mislevy, R.J. (1994b). 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.

Neapolitan, R.E. (1990). *Probabilistic reasoning in expert systems: Theory and algorithms.* New York: Wiley.

Nilsson, N.J. (1986). Probabilistic logic. *Artificial Intelligence, 28,* 71-87.

Noetic Systems, Inc. (1991). ERGO [computer program]. Baltimore, MD: Author.

Norman, D.A. (1988). *The psychology of everyday things.* New York: Basic Books..

Ohlsson, S. (1987). Some principles of intelligent tutoring. In R.W. Lawler & M. Yazdani (Eds.), *Artificial intelligence and education* (Vol. 1, pp. 203-237). Norwood, NJ: Ablex.

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

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

Savage, L.J. (1961). The foundations of statistics reconsidered. In J. Neyman (Ed.), *Proceedings of the Fourth Berkeley Symposium of Mathematical Statistics and Probability,* Vol. I (pp. 575-586). Berkeley: University of California Press.

Schum, D.A. (1979). A review of a case against Blaise Pascal and his heirs. *Michigan Law Review,* **77,** 446-483.

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

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

Shachter, R.D., & Heckerman, D.E. (1987). Thinking backward for knowledge acquisition. *AI Magazine,* **8,** 55-61.

Shafer, G. (1976). *A mathematical theory of evidence.* Princeton: Princeton University Press.

Shafer, G. (1987). Probability judgment in artificial intelligence and expert systems. *Statistical Science,* **2,** 3-16.

Shafer, G. (1988) The construction of probability arguments. In P. Tillers & E.D. Green (eds.), *Probability and inference in the law of evidence* (pp. 185-204). Dordrecht, The Netherlands.

Spiegelhalter, D.J. (1987). Probabilistic expert systems in medicine: Practical issues in handling uncertainty. *Statistical Science,* **2,** 25-30.

Spiegelhalter, D.J. (1989). A unified approach to imprecision and sensitivity of beliefs in expert systems. In L.N. Kanal, J. Lemmer, & T.S. Levitt (Eds.), *Artificial intelligence and statistics* (pp. 47-68). Amsterdam: North-Holland.

Spiegelhalter, D.J., & Cowell, R.G. (1992). Learning in probabilistic expert systems. In J.M. Bernardo, J.O. Berger, A.P. Dawid, & A.F.M. Smith (Eds.), *Bayesian Statistics 4* (pp. 447-465). Oxford, U.K.: Oxford University Press.

Spiegelhalter, D.J., Dawid, A.P., Lauritzen, S.L., & Cowell, R.G. (1993). Bayesian analysis in expert systems. *Statistical Science,* **8,** 219-283.

Szolovits, P., & Pauker, S.G. (1978). Categorical and probabilistic reasoning in medical diagnosis. *Artificial Intelligence*, **11,** 115-144.

Twining, W.L. (1985). *Theories of evidence: Bentham and Wigmore.* Stanford, CA: Stanford University Press.

Wenger, E. (1987). *Artificial intelligence and tutoring systems.* Los Altos, CA: Morgan Kaufman.

White, B.Y., & Frederiksen, J.R. (1987). Qualitative models and intelligent learning environments. In R. Lawler & M. Yazdani (Eds.), *AI and education*. New York: Ablex.

Wigmore, J.H. (1937). *The science of judicial proof* (3rd Ed.). Boston: Little, Brown, & Co.

Zadeh, L.A. (1965). Fuzzy sets. *Information and Control,* **8**, 338-353.

List of Tables

List of Figures