

Evaluation of an assessment system based on Bayesian student modeling

Kurt VanLehn

e-mail: vanlehn@cs.pitt.edu

Learning Research and Development Center

University of Pittsburgh,

Pittsburgh, PA 15260

Joel Martin

e-mail: joel@ai.iit.nrc.ca

Knowledge Systems Laboratory

National Research Council of Canada

Ottawa, CA K1A 0R6

Abstract. Schools need assessments of students in order to make informed decisions. The most common assessments are tests consisting of questions or problems that can be answered in under a minute each. When schools change their instruction to maximize performance on short-item tests, the students' learning can suffer. To prevent this, assessments are being developed such that "teaching to the test" will actually improve instruction. Such performance assessments, as they are called, have students work on complex, intrinsically valuable, authentic tasks. Olae is a performance assessment for Newtonian physics. It is based on student modeling, a technology developed for intelligent tutoring systems. Students solve traditional problems as well as tasks developed by cognitive psychologists for measuring expertise. Students work on a computer, which records all their work as well as their answers. This record is analyzed to form a model of the student's physics knowledge that accounts for the students' actions. The model is fine-grained, in that it can report the probability of mastery of each of 290 pieces of physics knowledge. These features make Olae a rather unusual assessment instrument, so it is not immediately obvious how to evaluate it, because standard evaluations methods assume the assessment is a short-item test. This paper describes Olae (focusing on parts of it that have not been described previously), several methods for evaluating complex assessments based on student modeling such as Olae, and some preliminary results of applying these methods to Olae with a small sample of physics students. In many cases, more data would be required in order to adequately assess Olae, so this

paper should be viewed more as a methodological contribution than as a definitive evaluation.

INTRODUCTION

Assessment is an important function of schooling. An assessment (test) is a decision aid. It produces information that allows students, teachers, parents, administrators and others to make better pedagogical decisions, such as which class to place a student in, whether to allow a student to go on to study the next unit in the syllabus, or which teacher should be nominated for national recognition. Assessment is not an end in itself, but serves only to improve decisions, inferences and actions.

Evaluation is the process of determining the value (worth) of an assessment. Since an assessment's role is to provide information to decisions makers, the only way to understand an assessment's value is to consider how it affects those decisions. At one time, tests were seen as measures of an underlying construct, so an evaluation was seen as determining the validity (truth) of the assessment. Nowadays, assessments are seen as one component of a decision making system, so they should be evaluated by their contribution to the system's overall performance. As the lead sentence of Messick's definitive article put it, "Validity is an integrative evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment." (Messick, 1989, emphasis in original).

Nowadays it is common to distinguish *evidential* from *consequential* validity. Evidential validity judges whether inferences and actions based *directly* on the assessment are adequate and appropriate. For instance, if one claims that a test predicts a student's aptitude for learning mathematics, then studies that correlate test scores with subsequent mathematical learning rates would evaluate the evidential validity of the test. In evidential assessment, one treats the test's stated purpose as a hypothesis, then tries to produce *evidence* that the hypothesis is correct.

Consequential validity judges the *indirect* consequences of using the test on the overall educational system. For instance, schools and teachers often change their instruction in order to improve students' performance on standardized tests. This is not necessarily bad. However, if the test assesses a deep, complex competence (e.g., verbal reasoning skill) with superficial tasks (e.g., vocabulary tests) that were once correlated with the complex competence, then "teaching to the test" means teaching students the superficial tasks instead of the complex competence. Because one consequence of using the test would be "dumbing down" the curriculum,

this test would have low consequential validity. On the other hand, if the assessment used tasks that really did tap the target competence, then teaching to the test would cause the assessment to have high consequential validity.

In order to improve both evidential and consequential validity, investigators are developing methods for directly assessing complex competencies based on extended student performances. Examples include open-ended mathematical problems, essays, hands-on science problems, computer simulations of real-world problems and portfolios of a student's best work. Such assessments were often referred to as "authentic" assessments (Linn, Baker, & Dunbar, 1991) because they involve tasks that have intrinsic value in themselves, rather than as correlates or indicators of valued performances. However, the term "performance assessment" seems to have become more common (Gall, Borg, & Gall, 1996).

The Olae system is a performance assessment. (The literature uses "assessment" to refer both to the tool or method that produces a report on the student's competence as well as for the report itself.) The students solve a variety of complex problem on a computer. The computer is almost as passive as piece of paper. It records the students' writing actions as they work, and sometimes where the students look. Olae analyzes these recordings and determines which pieces of domain knowledge and which learning strategies were used by the student.

"Olae" is an acronym for "On-Line Assessment of Expertise" because the students are using a computer (on-line) as they perform. It is also an acronym for "Off-Line Assessment of Expertise" because the student data are analyzed "off-line," that is, after the students have finished their work.

Olae is typical of other performance assessments in that it uses tasks that are complex and intrinsically valuable. However, Olae differs from other assessments in that it provides an unusually detailed report on the student's competence. For instance, it can report the probability of mastery of approximately 290 pieces of knowledge. Such a report is called a *student model* in the tutoring literature. Student models are used as diagnostic assessments by human tutors (e.g., Sleeman, Kelley, Martinak, Ward, & Moore, 1989) and by computer tutoring systems (VanLehn, 1988). In either case, they help the tutor make intelligent decisions about what pieces of knowledge to teach and how best to teach them.

The data analytic part of Olae has a formidable job. It must analyze second-by-second recordings of student performance on complex problems that can take an hour to solve, and it must produce a detailed report on the student's competencies (the student model). To handle this job, Olae employs a new technology called Bayesian networks (Pearl, 1988). Bayesian networks (also called belief networks, or causal networks, or

graphical models) makes it computationally feasible to use sound probabilistic reasoning about complex systems of relationships between data and results. Prior to their invention, only heuristic reasoning could be used with such systems (VanLehn, 1988).

The Olae system is only a prototype. It is intended to demonstrate that

1. it is possible to collect detailed performance data on student actions as they performed complex, intrinsically valuable tasks,
2. it is possible to analyze student competencies in detail, and
3. the data analysis can be sound and yet still computationally feasible.

These are primarily computational challenges. They have been achieved by demonstrating that Olae can be implemented and run with real student data (Martin & VanLehn, 1995a; Martin & VanLehn, 1995b). Moreover, the technology developed for Olae is now being used directly in two projects (Conati, Gertner, VanLehn, & Druzdzel, 1997a; Conati & VanLehn, 1995; Conati & VanLehn, 1996a; Conati & VanLehn, 1996b; VanLehn, 1996b) and similar technology is used in several others (Jameson, 1995).

Now that we have got a working assessment technology, it is appropriate to evaluate it. Evaluation of Olae is not easy, for three reasons.

First, Olae is a performance assessment. Standard methods of evaluation need extensive adaptation in order to apply to performance assessments (Linn et al., 1991; Messick, 1994). A national committee is revising the 1985 *Standards for Educational and Psychological Testing* to include performance assessment (Linn, 1994), but it is safe to say that more research is still needed.

Second, Olae produces a student model, which is a more detailed report on student competencies than other assessments produce. Many standard methods of evaluation assume that the assessment reports a single numerical or categorical value. Olae produces a large set of values.

Third, Olae's data analysis requires that many assumptions be made about the nature of competence in the domain. Although the data analysis algorithms are provably correct, Olae's assumptions about cognition are empirical claims and not subject to mathematical proof. Moreover, even if the assumptions are completely consistent with psychological evidence, this does not guarantee that an assessment built on top of them is valid.

Thus, it is difficult to evaluate Olae because it uses complex student performances, detailed reports of competence and multiple assumptions about cognition. Virtually any other performance assessment based on student modeling has these same problems. Nonetheless, designing a

performance assessment to be valid does not make it so. It is necessary to devise some means of evaluation for Olae and similar systems.

Several methods for evaluation are described herein. Although we applied them to Olae, we did not run the hundreds of subjects that would be needed for a real evaluation. Our intent is to show *how* Olae and other assessments based on student modeling could be evaluated. Indeed, we have uncovered several difficult technical problems that would have to be solved before a full-scale evaluation could be successfully completed. Thus, our major claim is that we have made progress toward finding methods for evaluating Olae and similar systems.

However, before describing the evaluation of Olae, we first describe Olae itself. The description of Olae varies in depth. It is shallow when describing parts of the system, such as the Bayesian data analysis, that have been covered in earlier reports (Martin & VanLehn, 1993; Martin & VanLehn, 1994; Martin & VanLehn, 1995a; Martin & VanLehn, 1995b). It is deeper when describing parts of the system that have been developed since then.

THE OLAE SYSTEM

Olae has been implemented with college physics as the task domain. Physics was chosen for several reasons. First, many physics educators feel that traditional assessment instruments overrate student's understanding in physics (e.g. Halloun & Hestenes, 1985; McCloskey, Caramazza, & Green, 1980). Second, physics has both procedural and conceptual content, which provides an interesting challenge for assessment. Third, solving a physics problem consists of a mixture of overt actions, such as drawing a vector or writing an equation, and covert reasoning that does not immediately result in visible actions, such as mentally envisioning the problem or planning a solution. The absence of overt action during important reasoning is typical of many authentic tasks, and makes performance assessment more difficult. Olae uses tasks invented by cognitive psychologists for measuring the differences between experts and novices, many of which were developed for physics, in order to supplement the data from problem solving and thus increase validity.

Olae has three components:

1. The task interfaces gather data from a student engaged in three traditional physics activities (solving quantitative problems, solving qualitative problems and studying worked examples) and three expert-novice tasks.
2. The data analyzers interpret data from the 6 task interfaces.

3. The assessor's interface displays the analyses graphically and at multiple levels of detail in order to facilitate informed decision making.

There are 6 data analyzers, one for each of the 6 tasks, but they all update a single data structure, the student model. The student model represents the probability of mastery of every piece of physics knowledge in the portion of physics covered by Olae. The student model will be discussed first, then each of the tasks and their data analyzers will be described. We conclude with a discussion of calibration, which is the process of finding values for numerical parameters required by the data analyzers.

We will not discuss the assessor's interface here (see Martin & VanLehn, 1995b, section 2.1). The purpose of the assessor's interface is to define and display aggregations of the fine-grained analyses. It is essentially an editor. It allows a human to graphically create a Bayesian network and attach it to the student model. For instance, if a human needs to make decisions about whether to advance a student after the student has studied chapter 5, and chapter 5 teaches a specific 11 rules, then the assessor can define a simple Bayesian network that computes the probability that the student has mastered all 11 rules in chapter 5 (e.g., $P(\text{Chapter-5-Mastery} \mid \text{Evidence}) = \prod P(\text{Rule-}k\text{-mastery} \mid \text{Evidence})$ for k from 1 to 11). The assessor's interface allows a human to define any function they want over the fine-grained assessment provided by Olae. After these functions are defined, they are computed every time the fine-grained assessments are computed. Thus, the human can define aggregate categories that are useful for the particular decisions need to be made. Because the assessor's interface is an editor, evaluating it would require real human assessors to solve real assessment problems with it. For evaluation purposes, we will ignore the existence of the assessor's interface and pretend that Olae generates only a student model.

The student model

Olae's representation of student knowledge consists of a set of rules, each expressing some small aspect of physics or algebra. For example, the following are 3 rules, which have been translated from their formal representation into English:

1. If there is a taut rope attached to an object, there is a tension force on the object.
2. If an object is moving with a constant speed, it has no acceleration.

3. If an object has no acceleration, the net force acting on it is zero.

The model is intended to contain all physics rules that a student could believe, even incorrect ones. (Rule 2 above is incorrect, for instance, because it assumes the object is moving in a straight trajectory.) Potentially, there are infinitely many incorrect rules that students could believe. This perennial problem with student modeling systems has received considerable attention (Baffes & Mooney, 1996a; Baffes & Mooney, 1996b; Burton, 1982; Kowalski & VanLehn, 1988; Langley & Ohlsson, 1984; Langley, Wogulis, & Ohlsson, 1990; Sleeman, Hirsh, Ellery, & Kim, 1990), and Olae includes no novel solutions to it. It simply includes rules for the most common misconceptions. Students with less common misconceptions will be misanalyzed, but this should happen infrequently. For the small fraction of physics that Olae covers (straight-line mechanics), 290 rules are used.

Olae uses a 3-level model of mastery. Each rule is assumed to be in one of 3 states:

1. Non mastery: The student never applies the rule.
2. Partial mastery: The student applies the rule when using paper and pencil, but does not use the rule when mentally planning a solution.
3. Full mastery: The student applies the rule whenever it is applicable.

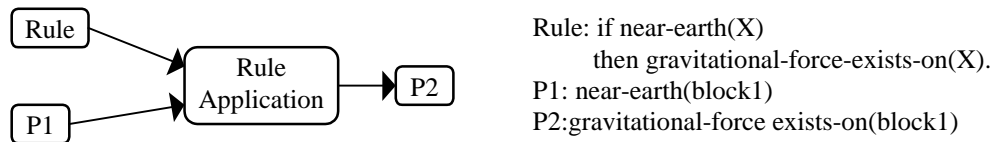
The distinction between partial and full mastery is motivated by a theory of expertise (Rubin, 1994; VanLehn, 1996a). The theory's claim is that experts are so familiar with the rules of physics that they can plan fairly detailed solutions in their heads without writing anything down. Intermediates' lack of familiarity with the rules prevents them from planning ahead. However, they do know the rules well enough to use them when aided by pencil and paper, because they can write the result of applying a rule down and do not have to hold them in memory while they recall the next rule and apply it.

Of course, even this 3-level scheme is a crude approximation. Whether a student can adequately remember a rule is contingent on many factors, such as how many times the rule has been accessed, how recently it was accessed and whether similar rules have been accessed recently (Anderson, 1983). Application is also contingent on the level of generality of the rule, or equivalently, how similar this application context is to the context in which the rule was learned. Moreover, a rule itself is not a single atomic item in memory, but a structure whose parts can be remembered or forgotten separately. In short, Olae's tri-state rules are a highly simplified

model of the student's knowledge. Nonetheless, it is much less of an approximation than a single number, such as "physics competence."

Given that Olae has 290 rules, each of which can be in one of 3 states (mastered, partially mastered and unmastered), there are 3^{290} possible states of knowledge that the student model can express. Most of the time, the system will be unable to determine exactly which state of knowledge best approximates the student's knowledge. For instance, if the system has not yet tested the student's knowledge of a particular rule, then it cannot know with certainty whether that rule is mastered, unmastered or partially mastered. However, it can guess. For instance, if students typically learn rules A and B at the same time, and there is evidence that the student knows A, then the system should infer that it is probable that the student knows B as well. In order to represent the system's uncertainty as to the student's state of knowledge and to allow probabilistic inferences based on trends in the student population, Olae should calculate the joint probability distribution over all states of knowledge. That is, for each of the 3^{290} states of knowledge, Olae should calculate the probability that the student's knowledge is best approximated by that state of knowledge.

Needless to say, storing 3^{290} numbers is impractical, so Olae uses a Bayesian network (Pearl, 1988; Russell & Norvig, 1995) instead of a tabular representation of the joint probability distribution. A Bayesian network is a directed graph whose nodes represent random variables. Each node has a number of states. The Cartesian product of the nodes' states is the set of states in the corresponding joint probability distribution. Thus, a Bayesian network with 10 binary nodes would represent a joint probability distribution over $2^{10}=1024$ states. A Bayesian network can represent any joint probability distribution and usually requires much less storage space than a tabular representation. Moreover, it makes important relationships such as independence immediately apparent, and there are many fast algorithms for doing calculations.



Prior Probabilities

Node “Rule”

P(unmastered) = .7
 P(partially mastered) = .2
 P(mastered) = .1

Node “P1”

P(in WM) = .95
 P(not in WM) = .05

Conditional Probabilities

“Rule Node Application”

| Rule state | P1 state | P(done) | P(not done) |
|--------------------|-----------|---------|-------------|
| unmastered | in WM | 0.0 | 1.0 |
| unmastered | not in WM | 0.0 | 1.0 |
| partially mastered | in WM | 1.0 | 0.0 |
| partially mastered | not in WM | 0.0 | 1.0 |
| mastered | in WM | 1.0 | 0.0 |
| mastered | not in WM | 0.0 | 1.0 |

Node “P2”

| Rule Ap. state | P(in WM) | P(not in WM) |
|----------------|----------|--------------|
| done | .999 | .001 |
| not done | .01 | .99 |

Figure 1. A simple Bayesian network

Olae’s Bayesian networks have a node for each rule. The node has 3 states: unmastered, partially mastered and fully mastered.¹ Each state has a probability indicating the chance that the student’s knowledge of that rule is in that state. Other nodes are attached or detached as necessary in order to allow interpretation of the data from the user interfaces. Most of these temporary nodes represent rule applications and propositions about the particular problem being solved. A rule application node has two states: done and not done. A proposition also has two states: in working memory and not in working memory. Figure 1 is a simplified example of a small fragment of Olae’s network. It shows a rule node, two proposition nodes, and a rule application node. Each node has a small table of probabilities associated with it. If the node has no incoming links, then the table represents the prior probabilities of that node. If the node has incoming links, then the table represents the probability of the node conditioned on

¹ Actually, Olae does not use a single node with 3 values. It uses two nodes with 2 values each. One node is True if the rule is either mastered or partially mastered, and False otherwise. The second is True if the rule is mastered, and False otherwise. Originally, Olae did not distinguish between partial and full mastery, so it only used the first node. It was easier to keep those nodes and add a new node than to revise them to use 3 levels of mastery.

each combination of the possible states of the parent nodes. All these probabilities must be provided in advance as part of the design of the network (see the section on calibration, below), and do not change. The topology of the network and the contents of these tables represent assumptions about cognition.

Also associated with each node is a posterior probability distribution that shows the probabilities of each node state given the evidence observed so far. These probabilities do change as Olaf observes the student. In **Error! Reference source not found.**, the posterior probabilities shown are those that the network would calculate after the student had been observed to draw a gravitational force vector on block1, and thus must have proposition 2 in working memory. Notice that the probability of the rule node has changed significantly. According to the prior probability, the student had probably not mastered the rule. After the vector was drawn, the posterior probability distribution shows that the student has probably either partially or fully mastered the rule.

| | |
|------------------------------|--------------------------------|
| Node “Rule” | Node “Rule Application” |
| P(unmastered) = .024 | P(done) = .976 |
| P(partially mastered) = .651 | P(not done) = .024 |
| P(mastered) = .335 | |
| Node “P1” | Node “P2” |
| P(in WM) = .999 | P(in WM) = 1.0 |
| P(not in WM) = .001 | P(not in WM) = 0.0 |

Table 1. Posterior probabilities after P2 observed

In point of fact, Olaf’s implementation of the Bayesian network technology was not particularly fast, so its networks were drastically simplified in order to conduct the experiments reported later. Instead of having a node for each of the 290 rules, it had nodes for only the 25 most central physics rules, including Newton’s three laws, the kinematics equations and some common force laws. (The network also has the temporary nodes, so it is much larger than 25 nodes.) The remaining 265 rules included a large number of rules representing algebraic, geometric and common sense reasoning, as well as physics rules that are only relevant to a few problems. Although this simplification of the student model hurts Olaf’s accuracy, it makes evaluation feasible computationally. In recent work, the computational limitations have been reduced by reimplementing the Bayesian network technology in C++ and using new update algorithms, which allows us to update large networks in few seconds (Conati et al., 1997a).

The student activities

This section describes each of the physics activities as presented to the student, the data recorded by Olae as the student performs, and the way the data are analyzed to update the student model. Although the evaluation of Olae focused primarily on data from the first activity, quantitative problem solving, the other activities are important because they are sensitive to cognitive processing that is difficult or impossible to measure given only quantitative problem solving data. Moreover, earlier reports on Olae did not indicate how it processed data from these activities, since that capability was added more recently.

The quantitative problem solving activity

The first activity to be described is quantitative problem solving. Quantitative problems are traditional end-of-the-chapter physics problems that ask students to calculate physical quantities such as accelerations or tensions for simple mechanical systems consisting of inclined planes, pulleys, blocks and so on. The user interface for this activity is intended to monitor in an unobtrusive way the student's performance as the student solves such problems. The computer screen is divided into several windows (Figure 2). Along the top are icons for specific physics problems. The student selects a problem by clicking on its icon. The problem is displayed in the upper left window. It consists of a statement of what is known and what needs to be found as well as a diagram of the problem situation. Below the problem description is a copy of the diagram. Students can draw vectors and coordinate axes on this diagram. The students enter equations in the window on the right. They are told to type everything necessary to solve the problem including side calculations and scratch work. The system records every vector, axes or equation entered.

In analysis mode, Olae processes each problem's data separately. Starting with the first problem solved by the student, it processes that problem's data, updates the Bayesian network, then processes the next problem's data, and so on.

To analyze a problem's data, the system starts with the current student model, which is a Bayesian network representing the student's current levels of mastery. It attaches a new network that represents all possible inferences that students can draw about this problem given the rules in the knowledge base. (This network is generated in advance by a rule-based problem solver and stored in a file. See Martin and VanLehn, 1995b.) Figure 1 showed a small fragment of such a network. Once the new network has been attached, the system is ready to interpret the student's actions. A student's action consists of entering a vector, coordinate axes or


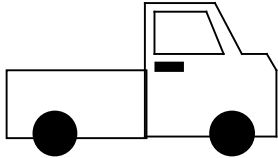
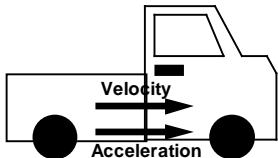
| Problem Solving | |
|---|--|
|  | |
|  A truck starts from rest and moves with a constant acceleration of 5 m/s^2 . Find its speed after 4 s. | $v = v + a * t$ $v = 0 + 5 * 4$ Answer: $v = 20 \text{ m/s}$ |
|  Velocity Acceleration | |
| Refresh | Examples |

Figure 2. The Olae screen during quantitative problem solving

equation. Olae figures out which node in the network each student's action denotes. Olae "clamps" that node to a state that represents that the node's proposition is known by the student. After the nodes for all the student's actions have been clamped, an algorithm is run that calculates, for every node in the network, the probability distribution across the node's states. **Error! Reference source not found.** illustrates how this calculation affects the probabilities of the nodes' states.

After the update, the network encodes the posterior joint probability distribution across the 3^{25} knowledge states (representing all possible combinations of mastery for the 25 target rules), given the collected evidence. One can easily read out of the network the marginal probability of mastery for each rule. With slightly more difficulty, one can also read out of the network the probability of any particular knowledge state. For instance, one can find out the probability that the student has partially mastered or mastered all the rules in a specified subset, such as the rules covered by Chapter 4. A tutor might find such assessments useful for deciding what type of problem to assign next to the student.

Usually students solve several problems with Olae, which means that several networks must be created and attached to the existing network of rule nodes, one for each problem. Since each of these problem-specific networks is quite large, the overall network can become too large to update in practical periods of time. Thus, Olae includes algorithms for compressing this huge network into a smaller one that includes only the rule nodes and a handful of extra nodes (Martin & VanLehn, 1994; Martin & VanLehn, 1995b). Compression causes some loss of information, so the

resulting smaller network only approximates the actual joint probability distribution of the original large network.

The example-studying activity

All textbooks contain examples of quantitative problems being solved. It has been found that students who study such examples by methodically explaining the example to themselves learn much more than students who merely read the example and paraphrase it (Bielaczyc, Pirolli, & Brown, 1995; Chi, 1996; Chi, Bassok, Lewis, Reimann, & Glaser, 1989; Chi, Leeuw, Chiu, & LaVancher, 1994; Ferguson-Hessler & Jong, 1990; Lovett, 1992; Pirolli & Bielaczyc, 1989; Renkl, 1997). In order to assess students' example studying strategy, Olae can present examples to students and monitor how they study them. It has also been found that when students self-explain a particular line of an example, they tend to learn rules involved in deriving that (VanLehn, submitted). Thus, a second goal for Olae is to infer, based on which lines the student self-explained, what the student probably learned from studying the example.

The example-studying user interface is similar to the quantitative problem solving user interface. The windows have the same information as in the quantitative problem solving interface, including vectors, axes and equations. However, the information is hidden until requested. Each equation in the right window is hidden by a shaded box. Boxes also hide the force diagram in the lower left window and the problem description in the upper left window. As the mouse arrow moves over a box, the box opens to reveal that part of the solution to the problem. The student can slowly step through the solution, opening one box at a time. This part of the example-studying interface is called the *poor man's eye-tracker*. It tells Olae what the student is reading and for how long. The example-studying interface can also be accessed during quantitative problem solving, but this capability is usually turned off, because we currently do not have a way of interpreting such example references.

Olae uses the duration of the student's study of a line in order to decide whether the student self-explained that line. Each line has a threshold which is set by hand. If the student studies a line for longer than the line's threshold, then Olae concludes that the student self-explained the line. A short look means self-explanation didn't occur.

If the student self-explained a line, Olae can assume that the student knows the rules involved in deriving that line. Most self-explanations consist of rederiving a line (Chi & VanLehn, 1991). Occasionally, this causes students to learn rules that they did not know before self-explaining the line (VanLehn, submitted). Regardless of whether a rule is known before self-explanation or learned during it, Olae should increase the

probability that the rule is mastered whenever a student self-explains a line whose derivation requires the application of that rule.

This simple interpretation of the eye-tracking data was motivated by a pilot experiment, wherein students gave verbal protocols as they studied an example. This allowed us to determine whether they were self-explaining a line or not. Self-explanation, as detected by verbal protocols, was indeed correlated with long latency (Martin & VanLehn, 1995a).

However, because there are many other cognitive processes that can cause long latencies, such as unsuccessful attempts at self-explanation or day dreaming, we did not feel confident enough in Olae's simple interpretation of the latency data, and thus did not completely implement it. In subsequent work, latency data were supplemented with a user interface that allows the students to express their self-explanations in a machine-readable form (Conati, Larkin, & VanLehn, 1997b). This richer data should allow a more reliable assessment of their knowledge.

The qualitative problem solving activity

In the 1970's, physics educators were surprised to discover that even students who received top grades in their physics courses could not answer certain qualitative questions which are, to a physicist, even easier than the quantitative questions that the students were trained on (e.g., McCloskey et al., 1980). The qualitative questions, however, are selected to elicit common student misconceptions. For instance, many students believe that when a rock is thrown upwards, there is an upward force on it as it rises, and a downward force on it as it falls. Nowadays, mastery of qualitative physics reasoning is considered an important objective of physics courses. Most recent physics textbooks include both qualitative and quantitative exercises.

In order to assess students' qualitative physics knowledge, Olae gives them qualitative problems to solve. The interface simply presents a student with a multiple choice question and collects the student's choice. Our questions and answer choices were taken from a widely used test, the Force Concepts Inventory (Hestenes, Wells, & Swackhamer, 1992).

Researchers currently do not completely understand the relationship between quantitative physics knowledge (represented in Olae with rules) and the kinds of informal, qualitative knowledge used to answer these question. In earlier work (Ploetzner & VanLehn, in press), one model of the relationship was developed and tested with student data. In subsequent, unreported work, we took verbal protocols of students as they answered Olae's qualitative questions. We found that students not only used different knowledge than they do during quantitative physics reasoning, they use different types of reasoning as well. They use proofs by

contradiction, extrapolation, interpolation, qualitative process modeling, qualitative algebra, case-based reasoning, and several other kinds of reasoning that defy classification.

As a consequence of this pilot work, we decided not to develop a thorough model of qualitative physics reasoning. Instead, we created Bayesian nodes for qualitative beliefs (including misconceptions) and linked them directly to each possible answer of each question. The qualitative beliefs were then linked to the relevant formal physics rules.

The solution planning activity

A perennial problem with student modeling systems is that students sometimes do quite a bit of important thinking without making a single interface action. Typically such lacunae appear just after a problem has been read, but they can occur at any time. When students are asked to provide a verbal protocol as they solve problems, their speech during these action-less periods often indicates that they are planning a solution to the problem. Their success at doing so depends strongly on their physics competence. Since Olae is intended to measure competence, it should try to determine how much planning capability the student has.

A direct way to uncover a student's competence in solution planning is simply to give them a quantitative problem and ask them to plan its solution without writing anything down. Physics instructors can plan the whole solution to a physics problem in their heads (Chi, Feltovich, & Glaser, 1981; Larkin, 1983).

For the problem in Figure 3, an expert might say something like,

Since the problem gives me everything I need to get the forces on the block, I can probably apply Newton's law to get its acceleration. But the problem doesn't want the acceleration. It wants to know how long the block will take to get to the bottom of the ramp. Since I know the ramp's length and the block's initial velocity, I can use kinematics to get the time from the acceleration.

On the other hand, when novices and intermediates are asked to find a plan for solving the same quantitative problems, their protocols are often quite incoherent and lack any evidence of a solution plan (Chi et al., 1981;

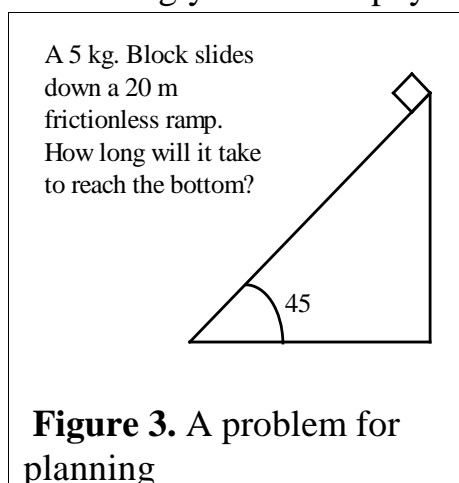


Figure 3. A problem for planning

Larkin, 1983). They say things about working methodically, scanning the textbook for useful equations, checking units and so on.

In order to measure the student's ability to plan the solution to a problem, Olae includes an computerized version of the Chi, Feltovich and Glaser (1981) solution planning task. Students are given a quantitative problem and asked to type in a statement of their basic approach to solving it. Their text is analyzed for keywords, such as "Newton's" and "kinematics," that tend to occur in coherent solution plans.

In order to interpret these data, it is useful to have a theory of how people plan solutions to physics problems. Jon Rubin developed a model based on the idea that planning consists of doing exactly the same sort of problem solving that one must do in paper-and-pencil solving, except that the equations are represented qualitatively as sets of equivalent quantities instead of algebraic expressions, and that minor equations are treated as mandatory substitutions (Rubin, 1994; VanLehn, 1996a). That is, a minor equation such as $F_{ax} = -F_a \cos(30^\circ)$, which is the projection of a force vector onto an axis, is never considered as a separate equation. Instead, wherever F_{ax} would appear in an equation, F_a is used instead. According to Rubin's theory, both experts and novices know how to plan with equations, but novices are not familiar enough with the requisite physics equations for them to do the substitutions mentally.

This account of solution planning was the main motivation for distinguishing partial mastery from complete mastery. A rule is said to be completely mastered if it can be used for both planning and paper-and-pencil problem solving. A rule is only partially mastered if it cannot be used during planning but can be used with paper and pencil.

The goal in analyzing the solution-planning data is to figure out which rules the person was using in their basic approach by using key words. If the student mentions the main principles for a problem but not the minor ones, then Olae increments the probability of full mastery on the rules that express the main principles and all the minor principles that are necessarily involved in this problem, even if the student does not mention the minor principles. If the student cannot come up with a plan, then the probability of full mastery for all the minor principles is lowered.

Difficulty estimation activity

Chi, Feltovich and Glaser (1981) found that expertise was correlated with the ability to give an accurate, well-founded estimate of the difficulty of a problem. Experts are not only able to more accurately estimate the relative difficulty of a set of problems, but they can provide more coherent, principled reasons for their ratings than novices.

One of Olae's activities is to have students rate the difficulty of a quantitative physics problem. The interface presents a problem, and asks them to rate the problem's difficulty numerically. It then asks them to list the factors that make the problem easy or difficult.

Although asking the students to rate the difficulty of the problem numerically is a necessary component of the overall activity, Olae does not actually analyze the students' numbers. The Chi et al. study showed that experts gave more accurate estimates of difficulty, in that the correlation with the actual time to solve the problem was higher than the equivalent correlation for novices. However, the difficulty estimate itself was not correlated with expertise (e.g., experts' estimates were not uniformly lower than novices'). This makes sense, given the theory of expertise mentioned earlier. Experts should be able to formulate a solution plan mentally and evaluate its difficulty, whereas novices should be unable to formulate a solution plan. However, the actual difficulty estimates given by the expert should not correlate with level of expertise because they are asked to predict the *relative* difficulty of problems. In short, there is no direct relationship between the students' estimate and their competence. As a consequence, Olae can not use the students' estimate. In the future, the interface could be modified to have the student solve the problem after estimating how long it will take them to solve it. Then the correlation between estimated and actual problem solving times could be taken as a gross indication of competence. However, the correlation would have to be measured across many problems, so it would be impossible to relate the correlation to knowledge of particular principles.

However, the difficulty estimation interface also asks students to identify the factors that make the problem easy or hard. This text is treated the same way as the text students enter to describe their basic approach to solving problems, which was described in the preceding section.

The problem classification activity

Olae's last activity has students sort quantitative physics problems into classes of similar problems. They can choose any definition of similarity they want. The user interface presents a problem and an icon for it. After reading the problem, the student drags the icon into a sorting area, and places it on the same line as the icons for similar problems. The student can easily change the existing classifications by dragging an icon from one line to another. Clicking on an icon displays its problem again.

Chi, Feltovich and Glaser (1981) discovered that experts usually define similarity to mean "the problems use the same basic physics principles in their solutions." Novices usually define similarity to mean "the problems

have similar physical objects and relationships.” Thus, a novice might group a pulley problem with another problem that has a pulley in it, whereas the expert might group it with a free-falling object problem because they both use conservation of energy in their solutions. This finding makes sense given the theory of expertise mentioned earlier. Because it takes expertise to plan a solution mentally, only experts can sort problems based on their solutions.

In order to analyze the classification produced by a student, each problem has a set of surface features and deep features assigned to it. The deep features are simply the rules used in solving the problem. The surface features encode objects mentioned in the problem statement (e.g., pulleys) and physical situations (e.g., free-fall). For each feature, Olae employs a Chi-squared test to determine whether the student used that feature in defining the classification. That is, if a feature present in almost all the problems in certain categories and almost always absent in the others, then it is probably being used in the classification. If it appears in the same proportion of the problems in each category, then it is probably not being used in the classification. The use of surface features indicates that the student has not reached complete mastery of the rules, so Olae lowers the probability of full mastery of all the rules used in the problems. Conversely, if the classification uses deep features (main principles), then Olae raises the probability of complete mastery.

Calibration

The calibration of Olae is the assignment of values to parameters so as to reflect frequencies in the student population. Population parameters in Olae include the prior joint probability distribution on rules, the conditional probability of a rule being applied given that it is mastered and all its inputs are known, the increase in probability of mastery indicated by use of a rule in classifying problems, and many others.

Ideally, one would take a large number of subjects, find out (somehow) exactly which rules they know, then have them perform Olae’s activities. This would allow us to set most of the parameters. For instance, one could set the prior joint probability distribution on the rules according to the distribution of rule mastery. One could also determine the conditional probability that a rule will be applied given that it is mastered. Because every parameter value is an estimate of the underlying value in the population, the larger the sample, the better the estimate.

Unfortunately, it is not practical to find out what rules hundreds of subjects know without using Olae itself. Verbal protocol analysis is the best known technique for uncovering rules, and it takes many hours of

analysis for each hour of protocol. Thus, we must use a less direct technique for calibrating Olae.

We devised a procedure for calibrating Olae, but we did not implement it as it would require assessing hundreds of subjects. The calibration procedure is based on the traditional EM (Estimation/Maximization) technique, which is a form of hill climbing.

For instance, consider the problem of setting of the prior joint probability distribution on rules. We assume that the structure of the Bayesian network for the rule nodes can be determined a priori. That is, we assume that it is easy to decide whether two rule nodes are linked based on considerations such as the sequence of topics in the curriculum and the logical compatibility of rules (e.g., a person is unlikely to believe both that the gravitational force is constant near earth and that it “wears off” as a object flies along). With this rather major assumption, the remaining task is to find values for the conditional probability tables of each node. Like all hill-climbing procedures, the first step of the EM procedure is to generate a random starting position. In this case, a random assignment of probabilities is generated. The next step is to run Olae on the subjects in the sample, and obtain their assessment. For each rule, we now know its frequency in the sample, according to Olae as calibrated with the random priors. The next step is to revise the prior probabilities to reflect the frequency distribution in the sample, then run Olae on each subject again. This gives us a new and presumably better estimate of the rules’ frequencies in the population. We again adjust the priors and run Olae. This continues until there is no significant change in the priors. From a hill-climbing perspective, we have reached a local maximum. We now repeat the whole process again, starting with a different random assignment of prior probabilities. This gets us to a new local maximum. We repeat the process many times, keeping track of the local maxima we reach, where each maximum is an assignment of prior probabilities. After many runs, we accept the most popular local maximum as a global maximum. In this fashion, we obtain the “best” setting of prior probabilities. In principle, the EM procedure can be used for setting all the parameters in Olae.

There is nothing particularly difficult about calibration, although it does use a great deal of computer time. The problem is that the parameter values obtained via calibration are good estimates of the population parameters only if the sample of students run on Olae is large. We feel that our subjective estimates are better than the parameter values we would obtain by performing the EM calibration procedure on the limited number of students that we have currently run.

In the next section, we report quantitative evaluations of Olae as well as qualitative ones. The quantitative evaluations use only the quantitative

problem solving activity. We believe that the quantitative problem solving data analysis is less sensitive to its parameters than the data analysis of the other activities, so the lack of calibration should not hurt it as much. In fact, we used a highly simplified parameter structure and still Olae was evaluated favorably. The parameter's structure was simplified in several ways. First, we assumed that all the nodes in the networks that relate rules to user interface actions are noisy-ANDs or noisy-ORs (see Pearl, 1988), so each node's conditional probability table is based on a single parameter that represents the amount of noise. Moreover, we used the same "noise" parameter value for all nodes, so there is only one parameter for those parts of the Bayesian network. Because the quantitative evaluations use only quantitative problem solving data, there is no way for Olae to differentiate partial mastery from full mastery, as the only activities that provide evidence for full mastery are solution planning, difficulty estimation and problem classification. Thus, for the quantitative evaluations, Olae's model of mastery was simplified to a binary value: non-mastery vs. partial or full mastery. For these evaluations, we assumed that all the rules had a prior probability of 0.5 of non-mastery and were conditionally independent of each other.

It bears repeating that in some cases, lack of data prevented us from carrying out the evaluation procedures described in the next section. Even though the evaluations that could be carried out all supported the validity of Olae, key ones are still missing, so the contributions of this paper are primarily methodological.

EVALUATION

Although the "worth" of an assessment is ultimately a unitary concept (it's either worth using or it's not), there are many different ways to decompose the concept (Messick, 1989). One decomposition, evidential vs. consequential validity, has been mentioned already. However, there are many others (e.g., Frederiksen & Collins, 1989; Linn et al., 1991) and no currently accepted standard. Before the acceptance of consequential validity, several national organizations developed a standard for evaluation of tests based on two fundamental concepts: reliability and validity (APA, AERA, & NCME, 1985). A test is reliable to the extent that it is free from random error (measurement error). A perfectly reliable test gives the same score every time it is applied to the same individual (assuming the individual is not affected by the test, which is rare). Reliability can be evaluated without knowing what the test score means. Validity, according to the 1985 standards, determines whether the meaning ascribed to the test

score is justified. Nowadays, that notion of validity would be called evidential validity.

In the absence of consensus on how to evaluate performance assessments (Linn, 1994; Linn et al., 1991; Messick, 1994), we will use a minimal extension of the 1985 standards. We will evaluate Olae in terms of its consequential validity, evidential validity and reliability.

In conclusion, calibration is important and can be activated via methods such as EM. Since we did not have enough data to apply such methods, we restricted our evaluation to the activity that we felt was least affected by the lack of calibration, and used a simplified parameter structure. Nonetheless, a “real” evaluation of Olae should be preceded by calibration.

Consequential validity

Olae has not been used in real schools yet, so we cannot say with certainty what its impact on them will be. However, we can estimate the impact based on the assumption that the educational system will “teach to the test.” That is, we can ask whether instruction would be hurt if it were changed to match the types of activities used by Olae.

Three Olae activities are currently used in instruction: quantitative problem solving, qualitative problem solving and example studying. Apparently, teachers think highly enough of these activities that they are willing to assign hours of work on them. Olae will probably not change that.

Olae’s other activities emphasize solution planning in one form or another. Dufresne et al. (1992) found that adding solution planning tasks to the instruction increased students’ learning. So this aspect of Olae probably increases consequential validity.

Olae does not monitor laboratory tasks, large-scale projects or other “hands-on” activities. It also does not monitor oral argumentation (“talking science”). If instructors decrease the amount of time devoted to these activities in order to devote more time to Olae’s activities, and these hands-on, talking-science activities are as instructionally valuable as they are thought to be, then consequential validity would suffer.

On the whole, Olae probably has higher consequential validity than traditional multiple choice or short answer exams. However, what one would really like is a more extensive version of Olae that monitors all the student’s work all the time (Collins, 1990). We will return to this point in the discussion section.

Evidential validity

Evidential validity views inferences based on an assessment as hypotheses. Evidential validity is the degree of support or evidence one can marshal for these hypotheses.

Before evaluating validity, one must clarify the hypothesis that one intends to support. Olae is intended to measure competence in physics, but there are clearly some parts of physics competence that are not addressed by any of its tasks, so it can't possibly assess them. For instance, it does not assess the students' skill at designing and conducting experiments, nor their skill at interpreting experimental data and arguing scientifically. All of Olae's tasks present a situation and ask for an analysis of it in one form or another. Thus, it is clear that Olae assesses a student's *analytical* competence, which is just one component of overall physics competence.

The 1985 standards define 3 basic types of validity. Messick (1989, pg. 16) paraphrases them as:

- “Content validity is evaluated by showing how well the content of the test samples the class of situations or subject matter about which conclusions are to be drawn.
- Criterion-related validity is evaluated by comparing the test scores with one or more external variables (called criteria) considered to provide a direct measure of the characteristic or behavior in question....
- Construct validity is evaluated by investigating what qualities a test measures, that is, by determining the degree to which certain explanatory concepts or constructs account for performance on the test.”

Content validity concerns the *authenticity* of the test: Does it tap all the important parts of the target competence? Criterion-related validity determines the correlation of the test score with some other measure of the target competence, if such a measure exists. For instance, because the SAT is designed to predict success in college, its scores should correlate with the students' college grade-point average. Construct validity determines whether the test score is confounded with personal characteristics other than the target competence. In the case of Olae, one should worry about whether its assessment is confounded with the subject's mathematical skill, their reading skill, their familiarity with computer interfaces, their typing skill or their familiarity with the specific physical systems (e.g., pulleys, springs) that happen to be used in the problems. Construct-validity also concerns whether the test suffers from racial, ethnic or gender bias. The

next 3 sections show how to apply these 3 types of validity to Olae, and when the data exist, the evaluations we conducted.

Content validity

Content validity asks, “Is the selection of test items representative of the ‘real’ domain tasks?” Content validity is often evaluated by asking teachers and other experts to rate the domain relevance of each item on a test. For Olae, a somewhat different method of evaluation seems warranted. However, it is still as judgmental and subjective as the standard method.

We discuss separately the knowledge tapped by a task and the type of problem solving it requires. For instance, a short answer test and a portfolio could in principle tap the same knowledge but require very different types of problem solving.

With regards to the type of problem solving, Olae’s quantitative and qualitative problems are the same format as those used in the instruction. In this respect, its content validity is arguably better than conventional testing that is based on short answer and multiple choice items. However, one could argue that even these long problems (they take about 15 or more minutes to solve) are not authentic enough. Students should be evaluated on multi-week, group research projects. Olae could play a small role in such authentic assessments by watching students do their calculations, but it cannot “listen in” to a group discussion, nor understand how much the teacher contributed.

In order to evaluate the knowledge employed by Olae’s tasks, we have the advantage of a detailed cognitive task analysis. For the quantitative problem solving activity, we know exactly what rules are necessary in order to provide a correct solution. For the other Olae activities, we have similar rule-level cognitive task analyses (Ploetzner & VanLehn, in press; Rubin, 1994), although those rules are not actually used in Olae. Moreover, these rule-based cognitive task analyses are computationally sufficient, in that a computer with no other source of knowledge than them can solve the tasks. Thus, these cognitive task analyses establish exactly what the content of the assessment is. The only question that remains is whether that content is representative of the task domain.

One way to understand the relationship of the tested knowledge to the overall competence is to reflect on the way the rule base changes as a new problem is added to Olae’s repertoire. When the new problem requires new physics principles for its solution, then significant changes are required that consist not only of adding new rules but sometimes changing old rules as well. This is to be expected, so in a production version of Olae, one would want to insure that all the major principles of physics were covered

by at least one Olae problem. However, a more interesting observation concerns the addition of new problems that, in the view of the physicist, should be solvable using the existing rules. We often find that such a problem cannot be solved because critical knowledge is missing. For instance, new rules are often needed to view the objects in the problem as ideal objects (particles, massless strings, frictionless surfaces, etc.), or to give a qualitative account of the object's motions. The inferences these rules make (e.g., that a block sliding down a frictionless plane will accelerate) are ones that are so obvious to physicists that they are not aware that they are making them. In short, even when the physicist thinks that the rules are sufficient to solve a new problem, it is often necessary to add a few minor, easily overlooked rules.

These observations bear upon the evaluation of content validity. If one had to add a great many such minor rules with the addition of every new problem, then the tested knowledge is probably only a tiny fraction of the overall domain knowledge. Moreover, possessing knowledge of some minor rules may not make it more likely that one possesses knowledge of others (except perhaps in the case of the mathematics rules). By keeping track of how many new rules are added as new problems are added, one might be able to fit a curve and make a well-founded estimate of the total number of rules in the domain. Moreover, studying the co-occurrence frequencies of minor rules would allow one to estimate the probability that the untested minor rules might be familiar already.

Unfortunately, we were not so methodical in developing Olae's rule base, so we have no such figures. We noticed that the number of rules added per problem dropped as we continued to add quantitative problems, so if we had to give an estimate, we think we might have to add another 100 rules to the existing 290 rules in order to cover all quantitative problems that one might find associated with Olae's target domain (straight line mechanics with Newton's law only; no curved trajectory, energy, work or momentum problems). Thus, for the knowledge that defines mastery of quantitative problem solving, content validity could be rather high because the untested rules are minor ones, there are probably not that many of them left (around 100) and many students might know them already. The content validity of the other activities remains unknown.

The most interesting outcome of this analysis is that it appears possible to evaluate content validity without relying as much on the judgments of experts. One uses the experts to generate a sample of problems that are representative of those found in the task domain as a whole. Then one implements a computational cognitive task analysis of the problems, keeping track of how many new rules are added with each problem. If the sample of problems generated by the experts is in fact a random one, then

one can extrapolate the curve of new rule additions to estimate the total number of rules in the task domain. A quantitative measure of content validity thus consists of the number of rules tapped by the assessment divided by the estimated total number of rules in the task domain.

Criteria-related validity

Criteria-related validity asks, “Do the test scores correlate with other measures of the target competence?” We are interested in concurrent validity only, since Olae is meant to assess the student’s analytical competence at the time of testing. The other form of criteria-related validity, predictive validity, is appropriate for assessments such as aptitude tests that predict the future performance of students. Evaluating the concurrent validity of an assessment involves administering another assessment of the target competence at approximately the same time as the assessment being evaluated.

We simulated the concurrent administration of two assessments by giving Olae raw data collected during an assessment conducted before Olae was developed. In the earlier assessment, a rule-based model of physics cognition, named Cascade, was fit to 9 subjects (VanLehn & Jones, 1993; VanLehn, Jones, & Chi, 1992). This fitting was based on verbal protocols as well as the worksheets of the subjects. However, the protocols included only quantitative problem solving, so that is the only Olae activity we could assess in this fashion.

We selected 2 of the more complex problems, because assessing problems that all subjects got right does not provide much of an evaluation for Olae. For each of the subjects, we entered into Olae the vectors and equations that the subject wrote on the worksheet, then had Olae assess the student. Thus, the human assessors had access to more data than Olae, because they had the verbal protocols as well as the worksheets. Thus, their assessment could be considered a “gold standard.”

The match between the human assessment and Olae’s assessment seems good. For each rule, the human assessor decided whether the student knew the rule or not. In all cases where the human assessors determined that the student knew a rule, Olae assigned a probability of greater than 0.85 to the rule. In the remaining cases, where the human assessors determined that the student did not know a rule, Olae assigned a probability of less than 0.15. That is, Olae and the human assessors were in 100% agreement.

In order to determine if this degree of match could occur by chance, we defined a null model as follows. We consider a “random” assessment that determines if a rule is mastered by flipping a weighted coin, when the coin’s weight depends on the rule being assessed, and the weight is the

proportion of students in the sample mastering that rule. For instance, if only 1 of the 9 students had mastered rule A, then the coin's weight for rule A would be 1/9. The observed degree of match between Olae and the human assessor was 100%. We want to calculate how likely it would be for the null model (tossing a weighted coin for each turn) to match the human assessor equally well, namely, 100%. For instance, suppose that Olae's assessment matched the human assessment on rule A for all 9 subjects. How likely is it that such a perfect match would be generated randomly, by replacing Olae with the weighted coin? For a student who had not mastered the rule, according to the human assessment, there is a 8/9 chance of the coin indicating non-mastery, and for the student who had mastered the rule, there was a 1/9 chance of the coin indicating mastery. Since there are 8 students who had not mastered the rule and one who had, the chance of the coin generating a perfect match to the human assessment is $(8/9)^8(1/9)=.043$. Thus, it is significantly unlikely ($p<.05$) that Olae's perfect match to the human assessment on Rule A mastery's is due to chance. On the other hand, suppose that all students had mastered rule B, and Olae also estimated that all students had mastered rule B. In this case, the weighted coin always picks mastery, so the chance of Olae matching the human assessors for all 9 subjects on rule B is $(9/9)^9=1.0$. That is, if all the students have mastered a particular rule or they have all failed to master a particular rule, then the fact that Olae correctly predicts the rule's mastery is quite unimpressive. In this fashion, we calculated that the probability is 0.000006 that the null model, a weighted coin, would agree as well with the human assessment as Olae did. Clearly, Olae's perfect agreement with the human assessors is no fluke.

These results indicate that Olae's criteria-related validity is quite high. However, only the quantitative problem solving activity was evaluated. Similar evaluations are needed for the other activities. They will probably not come out as well, because Olae's interpretation of that evidence seems to be based more strongly on parameters and Olae has not yet been calibrated. More importantly, we need to compare Olae's overall assessment, which includes data from all activities, to a similar one made by expert human judges from the same data. That would test not only Olae's interpretation of individual activities but its ability to integrate data from all the activities.

Construct validity

Construct validity, as used in the 1985 standards, is really a catch-all category that includes any method for testing the claims made by the assessment other than content validity or criterion-related validity. For instance, one method is to use protocol analysis to find out whether the

inferences that students find easy or difficult are mostly part of the target competence, or part of some other competence, such as mathematical reasoning or computer interface reasoning. Another would be to measure ethnic bias by testing large numbers of students from different ethnicity's. Another would be to see if reading or typing skill explained some of the variance in scores by administering tests for those subskills along with the test being evaluated. Clearly, there are so many methods for evaluating construct validity that one must choose only those that are most likely to reveal defects in the assessment.

One of the unusual aspects of Olae is its complex scoring method. This makes one wonder if it could be a source of invalidity. It may be that some pieces of knowledge are used so rarely that the Bayesian calculations cannot easily estimate the student's mastery of them, or it may be that the prior probabilities have too much influence. In order to evaluate these potential sources of invalidity, we developed a variant on criterion-related validity. Instead of using another assessment as a gold standard to compare Olae to, we generated "artificial students" who solved problems on Olae. Because we constructed the students, we knew exactly which rules they had mastered. Because we only have rules for quantitative problem solving, that is the only activity we assessed.

We generated 20 simulated students. For each student, we randomly selected a subset of the 25 rules as the ones that the student had mastered. However, we weighted our selection by the mastery of rules in the 9-student sample mentioned earlier. Thus, if a rule was mastered by all 9 real students, then it was mastered in all 20 artificial students. This selection procedure avoided generating artificial students which had mastered unlikely combinations of rules.

Each simulated student solved Olae's problems and Olae formed an assessment of the simulated student. We already know the precise set of mastered rules that underlies the student. The Bayesian network represents a joint probability distribution, so it can calculate the probability of that particular set of rules given the evidence. A very good score would be $0.95^{25}=0.28$ and a chance score would be $0.5^{25}=0$. Olae's score was 0.08, which is approximately $.90^{25}$.

As a standard for comparison, we also used the coin-flipping assessment described earlier. We calculated the probability that it would generate the correct combination of rule masteries. Its score was 0.002 ($=.78^{25}$), which is surprisingly high. The high value is due to fact that most of the rules had been mastered by all 9 subjects, so the coin-flipping assessment was able to accurately assess mastery of those rules. Nonetheless, Olae outperformed the coin-flipping assessment by an order of magnitude, which suggests that its assessment of the artificial students is valid.

Reliability

Reliability is the extent to which an assessment is free from random errors of measurement. For instance, your bathroom scale is unreliable if you can weigh yourself, step off the scale, weigh yourself again and get a different weight the second time. In order to find how unreliable your scale is, you can weigh many different objects twice and calculate the correlation of the first measurement of each object with the second measurement of each object. This is exactly what test developers do to evaluate an assessment. They test students twice and calculate the correlation between the first and second test scores. Correlation coefficients above .90 are common for standardized tests (Gall et al., 1996). An assessment with a correlation coefficient that is less than .70 is generally not considered suitable for individual student evaluations (Feldt & Brennan, 1989).

If the assessment measures a cognitive competence (as opposed to a physical one, say), then one cannot simply test the student twice. The student may remember responses from the first test and use them on the second test. This would artificially inflate the reliability of the assessment. Therefore, cognitive test developers typically evaluate reliability by giving subject just one test but analyzing it as if it were two parallel tests given simultaneously. For instance, they often treat the even numbered items as one test and the odd numbered items as another test. The scores on these two “tests” are calculated separately and correlated across subjects.

It appears that we could use this standard method to evaluation Olae’s reliability. However, there are two difficulties. Both are caused by the fact that Olae reports a student model instead of a single score as most assessments do. Thus, these difficulties would affect any assessment based on student modeling.

The first difficulty lies in differing conceptions of competence. The standard method of measuring reliability assumes that the student has a true score, which is a number indicating the student’s underlying competence. Each application of a test generates an observed score that is the sum of the true score and measurement error. All items on the test are assumed to be sensitive to the true score, which justifies dividing the test into parallel forms (e.g., the even and odd numbered items). However, Olae and other assessments based on student modeling view competence as mastery of many different pieces of knowledge. Solving a problem taps only some of the pieces of knowledge, and different problems can tap quite different portions of the student’s knowledge. Thus, partitioning the test creates two tests that may tap entirely different pieces of knowledge. That is, the two tests would not be parallel, as assumed in the standard method. Non-parallel subsets are particularly likely if the original test has only a

small number of problems, as is the case on most performance assessments including Olae.

The second difficulty is that the standard method of measuring reliability assumes that the assessment reports a single score. This makes it possible to measure correlations. If a test actually reports several scores (e.g., reading, math and writing), then they are usually combined linearly to form a single composite score (Feldt & Brennan, 1989). Although the composite score is not meaningful, it suffices for measuring reliability. Olae does not generate a single score but a Bayesian network that represents a joint probability distribution across approximately 25 variables. In principle, one could form a single composite score by generating all 2^{25} numbers in the joint probability distribution and summing them. This would make the composite score sensitive to dependencies among the variables. Clearly this is impractical. Another approach is to sum just the 25 marginal probabilities. This is feasible, but loses information about dependencies. Moreover, either method introduces dubious compensatory relationships. For instance, if test A says that two pieces of knowledge have probabilities of .5 and .9 respectively, and test B says they have probabilities of .95 and .45, then the two tests disagree and the reliability should suffer. Yet the sum of the two probabilities is 1.4 for both tests, thus hiding the different conclusions of the two tests in the composite scores.

These difficulties are purely technical. It should be possible to revise the standard method so that it will adequately evaluate reliability of assessments based on student modeling. However, we decided to try a different approach entirely.

Predictive accuracy is an evaluative measure used with systems that induce models of data (Russell & Norvig, 1995). One divides the data into two parts, called the training data and the test data. The to-be-evaluated system induces a model from the training data. The model makes predictions about the testing data. Comparing these predictions to the testing data establishes the predictive accuracy of the data analyzer. Typically, this process is repeated with many different partitions of the data into training and test sets.

To evaluate the reliability of Olae, we generated a student model using data from all but one of the student's problem solving performances, then use the student model to predict the student's performance on the remaining problem. In particular, the student model predicts exactly which vectors and equations the student will write.

As our sample, we used 5 subjects from the (Chi et al., 1989) study and 12 problems. For each subject, we ran Olae 12 times, leaving out one of the 12 problems and thus basing the student model on the student's responses to only 11 of the 12 problems. Olae calculated which rules the

subject had mastered, and using those it predicted the probability of the student's response on the left-out problem. In particular, the probabilities of each entry (each equation or vector) made by the student were multiplied to form the probability of the overall response.² This is a rather stringent measure. For instance, if a student entered 8 equations, and Olae predicted that each would be written with 0.99 probability, then its prediction of that exact combination of equations is $0.99^8 = 0.92$. If it made similar predictions for all 12 problems of all 5 students (i.e., 60 problems), then its prediction of the exact combination of observations is $0.92^{60} = .008$. Nonetheless, the probability assigned by Olae to the students' response (i.e., the product over all students, all problems and all entries) was 0.90, which appears satisfyingly high. In particular, it means that no entry in any problem of any subject had a predicted probability of less than .90.

Unfortunately, there are no standards for predictive accuracy as there are for reliability. In the data analysis literature, the predictive accuracy of a modeling technique is always evaluated relative to the predictive accuracy of a competing modeling technique. If there are no competitors in the literature yet, then one is invented for comparison purposes. When such a baseline modeling technique is needed, a common one is the "modal" model. In the case of Olae, the modal model simply outputs the same student model regardless of the student performance that is given to it. Given a sample of students, one builds a student model whose rules have a probability of mastery equal to the observed frequency of that rule's mastery in the sample. That is, if 4 of the 5 students have mastered rule A, then rule A is given a probability of mastery of .8 in the modal student model. All students are "assessed" as having the same profile of rule mastery, the modal student model. That profile is used to predict behavior on all the problems.

The modal student model was surprisingly good at predicting responses, and achieved a predictive accuracy of 0.78. This high value is due to the fact that the underlying distribution of rule frequencies was quite skewed. That is, most students really did have about the same profile of rule mastery, so the modal student was not such a bad approximation of their competence. Moreover, the rules whose mastery did vary across students affected only a small number of student entries. Most equations

²As pointed out to us by Albert Corbett, if Olae assigned a probability of 1.0 to all possible entries, it would always predict the student's entries with 100% accuracy. However, the cognitive model built into Olae assumes that student's will only use one strategy to answer a problem. Thus, it cannot predict that all possible strategies will be used, and thus that all possible entries will be made. Although Olae does not "cheat" by over-predicting, this particular evaluation method is unable to detect over-predicting.

and vectors could be generated with the knowledge that was mastered by all students.

The surprising success of the modal student may be a peculiarity of the 5-student sample used, or it may be a basic property of using predictive accuracy to evaluate reliability. More research is needed to understand how to appropriately measure the reliability of performance assessments based on student modeling. Until then, we cannot say how reliable Olae is. Moreover, this initial study of reliability only used the quantitative problem solving activity. A full-scale study should use all the activities.

LESSONS LEARNED

Our hypothesis is that Olae, and perhaps other assessments based on student modeling, are feasible, valid and reliable. We have made substantial progress in supporting this hypothesis. Although more work is needed, we now know what that work should be and where some of the major unsolved problems lie. Here we list the major lessons that we have learned.

Were the student activities appropriate?

Olae uses 6 student activities. Assuming that its goal is to assess a student's analytical competence and not other forms of physics competence, are these activities appropriate?

Quantitative and qualitative problem solving are clearly appropriate. They bear directly on analytical competence. We only wish that we could monitor qualitative problem solving more closely. It is difficult to infer a student's qualitative reasoning from the problem's final answer.

The example studying activity seems potentially quite valuable not only for assessing physics competence, but for assessing students' example studying strategies as well. That is, do they tend to self-explain the examples or not? However, we were not satisfied with using latency alone in determining whether a student has self-explained a line. In more recent work, a user interface has been developed that lets students enter their self-explanations in a machine-readable format (Conati et al., 1997b).

The expert-novice activities consisted of stating a basic approach, estimating difficulty and classifying problems. In laboratory studies, these tasks have been found to correlate with expertise. Olae uses them to assess not just a general level of expertise, but the mastery of individual principles used in planning solutions. It is not clear whether these tasks provide enough information to do that. Moreover, the information actually used by Olae is degraded on the basic approach task and difficulty estimation task, because the student's response is processed via keyword

analysis instead of full-fledged natural language understanding. The Andes intelligent tutoring system (VanLehn, 1996b) uses a form-based user interface that will allow students to express their basic approach without using natural language. At any rate, a necessary step for future research is to evaluate the criteria-related validity of these activities, perhaps by comparing their assessments to those obtained via protocol analysis.

Was a fine-grained, detailed assessment appropriate?

Olae produces a fine-grained, detailed assessment consisting of a student model that reports the probability of mastery of around 290 rules. This is certainly an unusual feature of Olae compared to other assessments. Is it useful?

Clearly, knowing only whether a student has mastered a particular rule does not allow an assessor to make far reaching decisions. Such detail is useful only for short-range decisions, such as those that a tutor might make: deciding what kind of exercise to assign or how to explain a complex concept. Indeed, when such fine-grained information was provided to human tutors by the Debuggy diagnostic assessment system (Burton, 1982), teachers reported that they only used it in order to decide what exercises to use during remediation (VanLehn, 1990). On the other hand, providing even more information than rule mastery, such as a precise description of a student's misconceptions, does not improve the effectiveness of human remedial tutors (Sleeman et al., 1989). Moreover, human tutors seldom diagnose to the level of misconceptions (Chi, 1996; McArthur, Stasz, & Zmuidzinas, 1990; Putnam, 1987), although they can report levels of mastery of individual concepts and principles. Thus, rule mastery seems to be about the finest level of detail that human tutors can use. Thus, grain size used by Olae seems appropriate for helping human tutors or teachers acting as tutors.

For other purposes, such as grading a homework assignment or a final exam, one would have to calculate some kind of score that aggregates over the rule mastery levels reported in the student model. Defining such a calculation can be done with Olae's assessor's interface. Would such a score be more useful than percent-correct or some other score calculated directly from the raw data?

Basing an aggregate score on the student model has one immediate advantage: the score can be explained. If the student asks, "Why did I get such a low score?" the assessor can say, "Because you have not mastered these 10 principles." Such specific diagnostic feedback would promote student learning more than telling them which problems they answered incorrectly. Basing aggregate scores on student models may increase the consequential validity of the assessment in less direct ways as well.

Because students can see that their score depends on mastering individual pieces of knowledge, they would focus their studying on mastering them and not on diffuse “studying the textbook.” This ought to increase their learning as well. Lastly, if instructors notice that many students have not mastered the same pieces of knowledge, then they might change their instruction. This would also increase the consequential validity of the assessment.

During the development and evaluation of a conventional test, developers sometimes do an item analysis, which amounts to deciding which pieces of knowledge are required by each item in order to correctly answer it. All the advantages of item analysis apply to assessments based on student modeling. For instance, using a student model as the basis for scoring defines exactly what the score is composed of. This makes it easier to judge the content validity of the assessment. It also allows the test developer to choose problems so that each piece of knowledge is used multiple times, and no single piece of knowledge is used on every problem. This increases both the evidential validity and the reliability.

Was the cognitive modeling worth it?

If one can get some of the advantages of a student model by simply doing an informal item analysis, was it worth the effort to develop rule-based computer models of physics problem solving? Granted, Olae was not the main motivation for our development of cognitive models of quantitative problem solving and example studying (VanLehn et al., 1992), qualitative problem solving (Ploetzner & VanLehn, in press) and solution planning (Rubin, 1994). However, could we have done just as well without these cognitive modeling efforts?

It seems necessary to have a cognitive model for interpreting the student’s behavior while solving quantitative problems. The only alternative would be to have experts solve each problem in all possible ways and record explicit lines of reasoning for each step of their solution. Each line of reasoning mentions exactly the rules that the expert applied. Although producing such traces of one’s reasoning seems less difficult than building a rule-based system, it is not clear whether it can be done with enough consistency and attention to detail. For instance, suppose that one discovered while solving the 20th problem that a certain rule should really be divided into two rules. One would have to go back and revise the solutions to most of the preceding problems. Thus, it is not clear which is easier: building a rule-based system or a rule-based analysis of each problem. However, it is clear that the analysis is necessary, regardless of whether it was constructed by hand or by running a rule-based problem solver.

For the other student activities (qualitative problem solving, example studying, etc.), the benefit of cognitive modeling was less clear. A detailed item analysis would perhaps have done just as well since Olae did not actually use the rule-based models for interpreting these activities' performance data.

If we were to do the Olae project over again, we would first build the user interfaces for each of the activities then run subjects on them while collecting verbal protocols. (We ran the subjects first, using pencil-and-paper tasks, then built the user interfaces.) On the basis of the protocols, we would develop cognitive models for some of the activities. This procedure would greatly increase the evidential validity. For instance, it would allow us to detect the degree to which the user interface was interfering with the solving of problems. Messick (1989, pg. 17) advocates this evaluation method above all others:

Historically, primary emphasis in construct validation has been placed on internal and external test structures, that is, on patterns of relationships among item scores or between test scores and other measures. Probably even more illuminating of score meaning, however, are studies of performance differences over time, across groups and settings and in response to experimental treatments and manipulations. Possibly most illuminating of all are direct probes and modeling of the processes underlying test responses....

Messick later (pg. 53) mentions that protocol analysis and cognitive model are one method for directly probing and understanding the processes tapped by a test. In short, even when cognitive modeling is not strictly necessary for the operation of the performance assessment system, it may be justified anyway because it increases our understanding of what the test actually measures.

Was the Bayesian data analysis worth it?

We adopted the Bayesian networks approach on the rather naïve belief that if we used sound algorithms for data analysis, the assessment would be more trustworthy. During the development of Pola, a successor to Olae (Conati & VanLehn, 1995; Conati & VanLehn, 1996a; Conati & VanLehn, 1996b), we discovered several mistakes in the way Olae built the Bayesian networks. This taught us that although the algorithms are sound, the structure of the network encodes many assumptions that may or may not be correct. Moreover, the values of prior probabilities can in principle dramatically affect the networks' behavior. This means that using Bayesian networks does not in itself guarantee a valid or reliable assessment.

However, the use of a sound algorithm does remove one source of potential data analytic problems. When the results seem wrong, we know

to look at the network structures and parameters. This is an improvement over heuristic techniques, where everything is suspect.

A particular feature of Bayesian networks is that they allow one to assign non-uniform prior probabilities, which is why they are called *Bayesian* networks. This means that if two lines of reasoning can both explain a student's response, the one whose rules have higher prior probability will get most of the credit for explaining the response. Those rules will get a bigger boost to their posterior probabilities than the ones on the less likely line of reasoning. This is exactly what one wants for handling ambiguous assessments, where the evidence does not uniquely identify the set of knowledge pieces mastered by the student. For instance, it is frequently the case in subtraction that the same answer can be generated by several different incorrect procedures (Burton, 1982). Moreover, even after a 20 problem test, it is often the case that there are many incorrect procedures that are each consistent with the student's answers. If one incorrect procedure is much more common in the population than any of the others (a typical occurrence, by the way), then one would guess that it is most likely that this student produced that answer via that procedure. Thus, when one must guess among options left open by the evidence, using Bayesian reasoning is advisable, provided of course that one has accurate estimates of the prior probabilities.

As it turns out, ambiguous assessments were uncommon in Olae. Because Olae monitors quantitative problem solving so closely, recording not only the final answer but the intermediate steps as well, there were few occasions when a student response could be explained by more than one line of reasoning. Thus, Olae seldom had to "guess," which is probably why Olae worked so well despite the fact that it had not been calibrated. In retrospect we could have used any sound algorithm for data analysis in place of one that allows non-uniform prior probabilities. This conclusion only applies to the quantitative problem solving activity. The other activities appear to have substantial ambiguity. For them, Bayesian reasoning and accurate calibration seem essential.

Is Olae better than other assessments?

The bottom line is, of course, whether Olae is better than other assessments. We first discuss the use of Olae as part of a course, then as a high-stakes standardized test. As usual, our criteria for "better than" are consequential validity, evidential validity and reliability.

With respect to consequential validity, Olae seems clearly better than assessments currently in use. Many high school and college physics courses are taught via lectures, labs, recitation sections and homework. Assessments are based on exams, lab reports and homework. If Olae

replaced the exams, students would probably devote more time to problem solving and less to memorizing the textbook. This would probably increase their learning, because many studies of cognitive skill acquisition indicate that learning-by-doing is more effective than studying a text (VanLehn, 1996a). Moreover, since some of Olae's activities (e.g., the qualitative problems) use the same format as exam questions, Olae could still assess important non-problem-solving knowledge, thus encouraging students not to ignore the textbook completely. Thus, in the context of this kind of physics instruction, Olae would increase consequential validity compared to the short-item tests being used now. If the course also included significant instruction in experimental design, data interpretation and scientific argumentation, all of which Olae does not tap, then either Olae would have to be extended or additional assessments would have to be included.

With respect to evidential validity, the evaluations described earlier indicate that Olae's quantitative problem solving activity does in fact measure the students' mastery of analytic physics knowledge. Similar evaluations of the other activities need to be done. On the other hand, we have not evaluated Olae's sensitivity to non-physics skills, such as typing, reading or mathematical manipulation, nor its fairness with respect to gender and ethnicity. Thus there remains some threat of confounds. Our major fear is that mathematical skill is incorporated in some of Olae's assessments of physics mastery. However, Olae's evidential validity probably meets or exceeds that of the exams used in current physics courses. In particular, they probably also confound physics and mathematics competence. Although an important goal for future research would be to more thoroughly evaluate Olae's evidential validity and compare it to commercially developed tests of physics competence, a major technical issue is that Olae produces a student model rather than a single score. This complicates both its evaluation and the comparison of its validity to the validity of ordinary tests.

With respect to reliability, technical problems have prevented us from fully evaluating Olae. Because Olae produces a student model rather than a single score, it is unreasonable to apply correlation-based techniques for measuring reliability. On the other hand, using predictive accuracy as we did makes sense only when comparing two assessments. Our comparison of Olae to a modal student model really didn't tell us much. As a consequence, we cannot say how reliable Olae is compared to other assessments, although the absolute value of the predictive accuracy of was satisfyingly high.

How does Olae fare when compared to other performance assessments? The major problem with all performance assessments is that they can only sample a small portion of a student's competence. Because it

takes hours or days for a student to complete just one task of a performance assessment, the assessment is limited to using only a small number of tasks. This hurts the reliability of the assessment, and hence its evidential validity (Linn, 1994; Messick, 1994). It can even hurt consequential validity, which is supposed to be the strong point of performance assessments. As Linn, Baker and Dunbar (1991, pg. 17) point out, “We should not be satisfied, for example, if the introduction of a direct writing assessment led to great amounts of time being devoted to the preparation of brief compositions following a formula that works well in producing highly rated essays in a 20-minute time limit.” As another example, if students are assessed via a portfolio consisting of the student’s two best pieces of work done during the semester, it makes considerable difference, both for evidential and consequential validity, whether the teacher has the students work on just those two pieces of work for the whole semester or work on dozens of pieces of work. The essential problem is that even with a performance assessment, the tasks used in the assessment are still only a *sample* of the student’s performance. Sampling can cause schools to focus their instruction too narrowly, hurting consequential validity, and sampling only taps only a small portion of the student’s knowledge, hurting reliability and evidential validity.

The right solution, as Collins (1990) points out, is to completely integrate instruction and assessment. That is, all student performances are monitored and play a role in assessment. Olae and similar systems make this feasible. Indeed, they could provide more information than is currently obtained. For instance, if Olae monitored all the student’s homework, then it could infer not only what the student knows now, but how long it took the student to learn it, what kinds of learning strategies the student employed and many other important features of the student’s learning performance. But how realistic is it to use Olae-graded homework as the primary assessments in a course? Are there cultural or practical impediments? Could Olae be extended to handle them?

First, homework assignments are necessarily unsupervised, “open book” tests. A valid assessment requires knowing how much help the student received while working. By examining the latencies between problem solving actions, it might be possible to infer whether the student was recalling physics principles from memory, looking them up in the textbook, or copying them from a friend’s solution.

Second, one might object that this is just like basing the students’ grades on homework assignments, which has always been an option for physics instructors and is rarely used nonetheless. One difference is that assignments are graded by Olae instead of human instructors, which makes grading large numbers of problems both feasible, more accurate and more

objective. Secondly, Olae can include the kinds of short-answer questions that are often used on exams now.

Lastly, Olae cannot tell who is sitting at the keyboard. Students could cheat by asking a friend (a generous friend!) to do all the student's homework. This problem could also be solved by having students do some of their homework in a supervised setting, such as a computer lab. Keystroke latency profiles or other authentication techniques could determine whether the same student was at the keyboard in both authenticated and nonauthenticated homework sessions.

In short, it seems that computer-monitored homework might be both an optimal performance assessment and a feasible one. Olae is a step in that direction.

However, we have so far considered using Olae only as a component of a course. Course grades are typically not useful except to those familiar with the course, because one course's grades not necessarily comparable to another's. For high stakes decision making by those unfamiliar with the student's courses, such as college admissions decisions, standardized testing is used. Comparability is assured by using the same test for all students and by having the test administered by trusted agents under standardized conditions. A major problem with using performance assessments for such purposes is that testing sessions are fairly short compared to the time required for an authentic performance, so either the students must submit portfolios of work completed outside the testing session or the performance assessment must compromise its authenticity by using simplified tasks. Olae and similar systems seem to be subject to the same constraint. Although the latency techniques mentioned above might be good enough for preventing cheating on homework, they are probably not sufficiently powerful to detect cheating on a high-stakes test that is administered outside a supervised location. Thus, the best way to use Olae or other performance assessments for high-stakes testing is to lengthen the testing sessions to days or weeks.

The bottom line seems to be that Olae and similar computer-monitored performance assessments are probably better than traditional course assessments, and if they were extended to monitor all homework assignments, they would certainly be better than current assessments. However, it appears that they are no more suitable than other performance assessments for high stakes testing. Basically, authentic performances by their very nature take a long time, and supervised testing sessions need to be short.

Acknowledgments

This research was sponsored by the Cognitive Science Division of the Office of Naval Research under grant N00014-91-J-1532. This paper was completed while the first author was a Fellow at the Center for Advanced Study in the Behavioral Sciences and supported by Spencer Foundation Grant number 199400132. The authors thank Abigail Gertner, Cristina Conati and Albert Corbett for reading the manuscript and commenting on it.

References

- Anderson, J. R. (1983). *The Architecture of Cognition*. Cambridge, MA: Harvard University Press.
- APA, AERA, & NCME. (1985). *Standards for Educational and Psychological Testing*. Washington, DC: American Psychological Association.
- Baffes, P., & Mooney, R. (1996a). Refinement-based student modeling and automated bug library construction. *Journal of Artificial Intelligence in Education*, 7(1), 75-116.
- Baffes, P., & Mooney, R. J. (1996b,). A novel application of theory refinement to student modeling. *Proceedings of the Thirteenth National Conference on Artificial Intelligence*, Portland, OR.
- Bielaczyc, K., Pirolli, P., & Brown, A. L. (1995). Training in self-explanation and self-regulation strategies: Investigating the effects of knowledge acquisition activities on problem-solving. *Cognition and Instruction*, 13(2), 221-252.
- Burton, R. B. (1982). Diagnosing bugs in a simple procedural skill. In D. Sleeman & J. S. Brown (Eds.), *Intelligent Tutoring Systems*. London: Academic Press.
- Chi, M. T. H. (1996). Constructing self-explanations and scaffolded explanations in tutoring. *Applied Cognitive Psychology*, 10, S33-S49.
- Chi, M. T. H., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science*, 15, 145-182.
- Chi, M. T. H., Feltovich, P., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, 5, 121-152.
- Chi, M. T. H., Leeuw, N. d., Chiu, M. H., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science*, 18, 439-477.
- Chi, M. T. H., & VanLehn, K. (1991). The content of physics self-explanations. *The Journal of the Learning Sciences*, 1, 69-105.

- Collins, A. (1990). Reformulating testing to measure learning and thinking. In N. Frederiksen, R. Glaser, A. Lesgold, & M. G. Shafto (Eds.), *Diagnostic Monitoring of Skill and Knowledge Acquisition*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Conati, C., Gertner, A., VanLehn, K., & Druzdzel, M. (1997a,). On-line student modeling for coached problem solving using Bayesian networks. *Proceedings of the 1997 User Modeling Conference, Sardinia*.
- Conati, C., Larkin, J., & VanLehn, K. (1997b). A computer framework to support self-explanation, *Proceedings of the Eighth World Conference of Artificial Intelligence in Education*.
- Conati, C., & VanLehn, K. (1995). A student modeling technique for problem solving in domains with large solution spaces. In J. Greer (Ed.), *Proceedings of the 1995 Artificial Intelligence and Education Conference*. Charlottesville, NC: Association for the Advancement of Computers in Education.
- Conati, C., & VanLehn, K. (1996a). POLA: A student modeling framework for probabilistic on-line assessment of problem solving performance. In D. N. Chin, M. Crosby, S. Carberry, & I. Zukerman (Eds.), *Proceedings of UM-96, the Fifth International Conference on User Modeling* (pp. 75-82). Kailua-Kona, Hawaii: User Modeling, Inc.
- Conati, C., & VanLehn, K. (1996b). Probabilistic plan recognition for cognitive apprenticeship. In J. Moore & J. Fain-Lehman (Eds.), *Proceedings of the Eighteenth Annual Meeting of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.
- Dufresne, R. J., Gerace, W. J., Hardiman, P. T., & Mestre, J. P. (1992). Constraining novices to perform expert-like problem analyses: Effects on schema acquisition. *The Journal of the Learning Sciences*, 2(3), 307-331.
- Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational Measurement* (Third ed., pp. 105-146). New York: Macmillan.
- Ferguson-Hessler, M. G. M., & Jong, T. d. (1990). Studying physics texts: Differences in study processes between good and poor solvers. *Cognition and Instruction*, 7, 41-54.
- Frederiksen, J. R., & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher*, 18(9), 27-32.
- Gall, M. D., Borg, W. R., & Gall, J. P. (1996). *Educational Research: An Introduction*. (6th ed.). White Plains, NY: Longman.
- Halloun, I. A., & Hestenes, D. (1985). Common sense concepts about motion. *American Journal of Physics*, 53(11), 1056-1065.

- Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *The Physics Teacher*, 30, 141-158.
- Jameson, A. (1995). Numerical uncertainty management in user and student modeling: An overview of systems and issues. *User Modeling and User-Adapted Interaction*, 5.
- Kowalski, B., & VanLehn, K. (1988). Cirrus: Inducing subject models from protocol data. In V. Patel (Ed.), *Proceedings of the Tenth Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Langley, P., & Ohlsson, S. (1984,). Automated cognitive modeling. *Proceedings of the National Conference on Artificial Intelligence*, Austin, TX.
- Langley, P., Wogulis, J., & Ohlsson, S. (1990). Rules and principles in cognitive diagnosis. In N. Frederiksen, R. Glaser, A. Iesgold, & M. Shafto (Eds.), *Diagnostic Monitoring of Skill and Knowledge Acquisition* (pp. 217-250). Hillsdale, NJ: Erlbaum.
- Larkin, J. (1983). The role of problem representation in physics. In D. Gentner & A. Stevens (Eds.), *Mental Models*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Linn, R. L. (1994). Performance assessment: Policy promises and technical measurement standards. *Educational Researcher*, 23(9), 4-14.
- Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher*, 20(8), 15-21.
- Lovett, M. C. (1992). Learning by problem solving versus by examples: The benefits of generating and receiving information. *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Martin, J., & VanLehn, K. (1993). OLAE: Progress toward a multi-activity, Bayesian student modeller. In S. P. Brna, S. Ohlsson, & H. Pain (Eds.), *Artificial Intelligence in Education, 1993: Proceedings of AI-ED 93*. Charlottesville, VA: Association for the Advancement of Computing in Education.
- Martin, J., & VanLehn, K. (1994). Discrete factor analysis: Learning hidden variables in Bayesian networks. Technical report: LRDC, University of Pittsburgh.
- Martin, J., & VanLehn, K. (1995a). A Bayesian approach to cognitive assessment. In P. Nichols, S. Chipman, & S. Brennan (Eds.), *Cognitively diagnostic assessment*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Martin, J., & VanLehn, K. (1995b). Student assessment using Bayesian nets. *International Journal of Human-Computer Studies*, 42, 575-591.

- McArthur, D., Stasz, C., & Zmuidzinas, M. (1990). Tutoring techniques in algebra. *Cognition and Instruction*, 7(3), 197-244.
- McCloskey, M., Caramazza, A., & Green, B. (1980). Curvilinear motion in the absence of external forces: Naive beliefs about the motion of objects. *Science*, 210(5), 1139-1141.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3 ed., pp. 13-103). New York: Macmillan.
- Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, 23(2), 13-23.
- Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo, CA: Morgan-Kaufmann.
- Pirolli, P., & Bielaczyc, K. (1989). Empirical analyses of self-explanation and transfer in learning to program. *Proceedings of the Eleventh Annual Conference of the Cognitive Science Society* (pp. 459-457). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Ploetzner, R., & VanLehn, K. (in press). The acquisition of informal physics knowledge during formal physics training. *Cognition and Instruction*.
- Putnam, R. T. (1987). Structuring and Adjusting Content for Students: A study of Live and Simulated Tutoring of Addition. *American Educational Research Journal*, 24(1), 13-48.
- Renkl, A. (1997). Learning from worked-examples: A study on individual differences. *Cognitive Science*, 21(1), 1-29.
- Rubin, J. (1994). *A model of expert problem solving in elementary mechanics*. Unpublished Masters, Intelligent Systems Program, University of Pittsburgh, Pittsburgh, PA.
- Russell, S., & Norvig, P. (1995). *Artificial Intelligence: A Modern Approach*. Los Altos, CA: Morgan-Kaufmann.
- Sleeman, D., Hirsh, H., Ellery, I., & Kim, I. (1990). Extending domain theories: Two case studies in student modeling. *Machine Learning*, 5, 11-37.
- Sleeman, D., Kelley, A. E., Martinak, R., Ward, R. D., & Moore, J. L. (1989). Studies of diagnosis and remediation with high school algebra students. *Cognitive Science*, 13, 551-568.
- VanLehn, K. (1988). Student modeling. In M. Polson & J. Richardson (Eds.), *Foundations of Intelligent Tutoring Systems* (pp. 55-78). Hillsdale, NJ: Lawrence Erlbaum Associates.
- VanLehn, K. (1990). *Mind Bugs: The Origins of Procedural Misconceptions*. Cambridge, MA: MIT Press.

- VanLehn, K. (1996a). Cognitive skill acquisition. In J. Spence, J. Darly, & D. J. Foss (Eds.), *Annual Review of Psychology*, Vol. 47 (pp. 513-539). Palo Alto, CA: Annual Reviews.
- VanLehn, K. (1996b). Conceptual and meta learning during coached problem solving. In C. Frasson, G. Gauthier, & A. Lesgold (Eds.), *ITS96: Proceeding of the Third International conference on Intelligent Tutoring Systems*. New York: Springer-Verlag.
- VanLehn, K. (submitted). Rule learning events in the acquisition of a complex skill: An evaluation of Cascade.
- VanLehn, K., & Jones, R. M. (1993). Learning by explaining examples to oneself: A computational model. In S. Chipman & A. Meyrowitz (Eds.), *Cognitive Models of Complex Learning* (pp. 25-82). Boston, MA: Kluwer Academic Publishers.
- VanLehn, K., Jones, R. M., & Chi, M. T. H. (1992). A model of the self-explanation effect. *The Journal of the Learning Sciences*, 2(1), 1-59.