AUTOMATED PERFORMANCE ASSESSMENT TOOLS

John Leddo, Ph.D.
Research Development Corporation
Herndon, VA 20171-3205

Zhixiong Zhang
Research Development Corporation
Herndon, VA 20171-3205

Robert Pokorny, Ph.D.
U.S. Air Force Research Laboratory
Brooks AFB, TX 78335-5352

Measuring performance, whether to assess the results of training or select personnel for promotion, has never been more important. Organizations have generally been successful at defining criteria for satisfactory performance. Developing ways of measuring performance that are valid, reliable and economical to administer have been more problematic. Human evaluators are often costly and have low inter intra-/rater reliability. While automated assessment tools may be cost-effective and consistent across ratings, they tend to be limited in their understanding of the domain they are assessing. Hence, the quality of their assessments has been questioned.

Research Development Corporation has developed PC-based automated performance assessment technology.   The testbed is in-patient care provided by medical technicians. The technician being assessed performs required tasks in a simulated environment and is assessed according to criteria established in the Air Force Career Field Education and Training Plan (CFETP). The system uses both the technician's behaviors during the simulation and responses to questions presented by the tool to assess performance. The tool outputs a score that is based on the CFETP scoring system and an explanation of the score to support training.

The technology seeks to overcome limitations of other scoring systems by incorporating an expert model of the task being performed. The system represents the knowledge using RDC's integrated knowledge structure (INKS) framework that contains knowledge of causal principles, goal and planning knowledge, procedures and factual knowledge (which correspond to the knowledge types outlined in the CFETP). The tool runs the technician's behaviors through its expert model to determine whether the technician's solution meets the task requirements. It follows up with questions based on the INKS knowledge types to insure that the technician not only can perform the task, but has the deeper understanding of its underlying concepts and principles.

ABOUT THE AUTHORS

Dr. John Leddo is President of Research Development Corporation. He holds a Ph.D. in Psychology from Yale University. Dr. Leddo specializes in developing training technologies along with the cognitive learning models upon which the technology is based.

Mr. Zhixiong Zhang is the lead software engineer at Research Development Corporation. He is a Ph.D. candidate in Information Technology at George Mason University. Mr. Zhang specializes in developing simulation-based intelligent tutoring systems.

Dr. Robert Pokorny is a Research Psychologist with the Air Force Research Laboratory. He holds a Ph.D. in experimental psychology from the University of Oregon. Dr. Pokorny specializes in developing methods for efficient knowledge acquisition for complex tasks.

INTRODUCTION

Today, the focus on assessing performance is as high as it has ever been. Both public and private organizations are developing measures of organizational performance that trickle down to the individual worker level. These measures are used anywhere from hiring to promotion and job selection decisions to determining training needs of the workforce to deciding whether to terminate someone for lack of performance.

Because of the high stakes nature of measuring performance, the validity of the measurement instrument is typically the focal point. Currently, there is a strong push to make assessments more “authentic”. “Authentic” means that the assessments need to measure knowledge and skills in a way that is more closely matched to how those skills are used in a practical setting. This is in response to more traditional multiple choice assessment methods that focus more on factual knowledge and the ability of test takers to memorize the content that will be tested. Such tests have been criticized as measuring a person’s ability to take a test rather than to really master the skills being tested.

In response to these criticisms, a new trend in authentic assessment has focused on performance tasks, where people are asked to solve practical problems. As people perform these tasks, a protocol is taken of their problem solving process and solution, typically in written format. This protocol is then content analyzed by independent raters who are provided a scoring key or rubric. Typically, the rubric is anywhere from a 3 to a 6 point scale where the highest score reflects the correct answer with sound reasoning, the lowest score reflects no answer or an incorrect answer with poor reasoning, and the middle scores reflect some combination of correct answer/poor reasoning or incorrect answer/good reasoning.

There are weaknesses to performance assessment that we address in our present work. The most critical is the lack of a framework that operationalizes the standards into the kinds of knowledge and reasoning skills necessary to master the topics. As a result, there is little consistency across scoring rubrics for performance tasks and the scoring rubrics themselves do not provide feedback as to why students have problems. A second weakness is that performance assessments typically do not provide an explanation of how the score was arrived at. A performance level that reports “partial mastery” (a common scoring category) does not say which skills have been mastered and which are lacking. Therefore, it is hard to use those results for remedial training.

We do not believe that the Performance Assessment paradigm is fundamentally flawed. However, it is still in its relevant infancy. We believe that a major problem with the performance assessment paradigm is the lack of a model of the problem solving knowledge and skills that support problem solving behavior. Such a model could be used to operationalize what different performance levels really are. Just as scientists take general questions and operationalize them into things they can investigate and measure, educators need a methodology for operationalizing their standards into things that can be taught and measured (assessed). Another key weakness (which forms the basis for criticism of the paradigm) is the subjectivity of the rating process, leading to potential biases based on who the raters happen to be.

Interestingly, the most intensive effort to address the issue of developing a knowledge model against which performance can be evaluated against has come in the field of intelligent tutoring systems (cf., Brna, Ohlsson and Pain, 1993; Greer, 1995). An intelligent tutoring system (ITS) is a training system that has an underlying knowledge model of both the domain and the student and seeks to instruct the student based on his or her performance against the domain (often called the expert) model.

The ITS’ expert model serves as the operationalization of student performance. Typically, the ITS works by presenting the student with a problem. Based on that problem, the ITS’ expert model has a set of specified solution paths, often represented in terms of acceptable procedures that can be followed, against which it compares the student. The student is then evaluated on whether his or her process matches a known problem solving protocol.

A variation of this is seen in ITSs that employ simulation/scenarios as the primary method of assessing problem solving skills (cf., Kraje, Woolf, Fray, and Ghosh, R., 1993; Leddo and Kolodziej, 1997). Here, a person is presented with a scenario depicting a problem that needs solving (e.g., a patient is having a heart attack). The person then enters the steps s/he would take in that situation. The ITS then does two things. First, it compares the person's actions to its own expert model to determine whether the solution is acceptable. Second, it updates the scenario to show the consequence of the action taken by the person. At this point the person is faced with having to solve the problem based on their initial actions (e.g., how the patient has responded to the treatment given by the student). The internal expert model generates a new set of acceptable solutions and the cycle iterates. The resulting feedback measures when and where the person deviated from the prescribed procedure path, taking into account the logical scenario changes induced by the person's previous actions.

The above paradigm represents interesting improvements to the performance assessment problems cited above. First, there is an underlying model that prescribes what the behaviors a person should take are. Second, the evaluation process occurs consistently across people. There are still weaknesses that need to be addressed further.

The most important weakness is the overreliance on procedural knowledge. This makes novel solutions difficult to evaluate. Related to this, not all problem solvers think in the type of linear mode that procedural problem solving suggests.

Another important weakness in this paradigm is the "all or none" nature of the evaluation. While it is true that the evaluation may point out areas in which the person did follow the expert model as well as areas that s/he did not, the fundamental evaluation of each skill component is "followed the expert model/did not follow it". While error catalogs are useful in helping hypothesize where mastery falls short and what remediation is needed, this procedural approach is geared more toward evaluating whether or not a procedure was followed and not at giving "partial credit" to help evaluate gradations of skill level for specific tasks.

Another weakness is that observance of behavior does not reflect true mastery of subject matter. Two people can exhibit the same behavior for different reasons. One can perform exactly as instructed and appear competent on a test of skill based on a program of instruction. However, this person may have little deep understanding of the underlying principles and therefore be unable to transfer or extend these skills in a novel problem. Another person may have little procedural knowledge, but can use general principles to derives procedures in almost any problem solving situation. Clearly the latter person has greater mastery of the domain, but without in-depth probing of the latter knowledge it is difficult to deduce this skill from the scenario behavior alone.

The present paper describes our work with the United States Air Force to develop automated performance measurement tools. In developing these tools, our goal was to address the weaknesses of the approaches reviewed above while preserving their strengths.

AUTOMATED PERFORMANCE ASSESSMENT TECHNOLOGY

Testbed Selection

The Air Force provided us a list of candidate occupations on which we could focus. We selected medicine as the profession to work with because of its dual value to both the Air Force and private sectors. Once the medical profession was selected, we established, with the help of our Air Force sponsor, proponency at Wilford Hall Medical Center (WHMC) in San Antonio, Texas.

With the help of our Air Force proponent, we tried to identify a promotion track of high priority for which performance could be measured. In conjunction with WHMC, we selected the occupation of Medical Service Technician (AFSC number 4N0X1) and focused on the requirements for promoting a Technician from a 3 level (apprentice) to a 5 level (journeyman) or from a 4N031 to a 4N051.

The Air Force Career Field Education and Training Plan for the Medical Service Specialty outlines the knowledge and skills personnel must have to be promoted to different levels. For example, in order to be promoted from a 3 level to a 5 level, a worker must demonstrate mastery of 5 level knowledge and skills. We worked with WHMC to select appropriate knowledge and skills for this promotion requirement that could be used in our technology. We decided to focus on inpatient services and, in particular, the high priority task of measuring and recording fluid intake and output.

There are many medical conditions in which the amount of fluid in a patient's body is of critical importance (e.g., congestive heart failure, dehydration). Because of this, part of the patient's care and monitoring centers around controlling and measuring the amount of fluid that goes in and out of a patient's body.

Some of the mechanisms by which a patient takes in or puts out fluid are more controlled and easily measured than others. For example, how much fluid a patient gets through an IV is typically under the control of the physician and can be measured using the IV apparatus itself. On the other hand, the patient can drink fluids orally. In such cases, the control is often the patient's and monitoring that type of fluid intake requires the cooperation and the competence of the patient.

Similar cases apply to fluid output. A patient can eliminate fluid (referred to as "voiding") through a catheter or through urination. The former is easier to monitor. Additionally, there are other avenues of fluid output such as sweating that are very difficult to monitor and measure.

As a result, there are standard procedures that are followed like checking the IV levels to see how much fluid a patient has taken in intravenously and monitoring the foley catheter to see how much fluid the patient has voided. There are less well structured aspects of the task as well such as determining if a patient has taken in or put out fluids in other ways (e.g., did relatives come for a birthday party and bring cake and milk, did the patient get hot flashes and sweat profusely?).

There are other interesting aspects of this task as well. Some of the information the technician receives has a high degree of certainty (e.g., IV fluid intake). Other information may have low certainty (e.g., a patient reports having "a couple of glasses of water"). A technician must deal with this uncertainty, such manner of dealing with it being largely affected by the patient's medical condition. For example, if a patient is on fluid restriction, it may be better to overestimate fluid intake when there is uncertainty. On the other hand, if a patient is dehydrated, it may be better to underestimate it.

Finally, the technician needs to consider health implications for when there is a deviation from desired fluid intake and output levels. For example, too little fluid intake may result in a drop in the patient's blood pressure. Technician's need to know these potential consequences and the signs that they are occurring.

Conduct Knowledge Engineering With Domain Experts To Build Knowledge Base For The Tool.

The next step in the project was to conduct knowledge engineering with WHMC-supplied medical experts to build the knowledge base that the assessment tool would use to measure medical technician performance. WHMC provided us with three subject matter experts.

Prior to performing the knowledge engineering sessions, the following preparation was done. As noted earlier, we had selected promotion from an AFSC 4N031 ("3 level") to an AFSC 4N051 ("5 level") as the focus of the project (or promotion of a medical service technician from an apprentice to a journeyman). The Air Force publishes a Career Field Education and Training Plan that outlines the knowledge and skill requirements for such a promotion to occur. In addition to outlining the requirements, there is a performance scoring system that specifies the level of proficiency required for each skill.

Because we knew in advance that we would be working with medical service technicians on the task of fluid intake and output, we reviewed the Career Field Education and Training Plan for medical service technicians to determine what the proficiency requirements were for the task of fluid intake and output to be promoted from 3 level to 5 level. According to this document, there are two categories of requirements: task performance levels and task knowledge levels.

The task performance level for promotion for this task is that the technician "Can do all parts of the task. Needs only a spot check of completed work." The task knowledge level required for promotion for this task is that the technician "Can identify why and when the task must be done and why each step is needed."

These requirements drove the topics covered in the knowledge engineering session. In particular, to cover the task performance level requirement, we queried the subject matter experts (SMEs) on the procedures involved in measuring and recording fluid intake and output. These included the equipment used in the hospital (e.g., IV's, foley catheters) and recording forms (for which we obtained copies). To cover the task knowledge level requirement, we queried SME's on the types of medical conditions that fluid intake and output is important to, the implications of different deviations from normal in a patient's intake and output and how to recognize them, how does a technician know when to perform the measurement and recording, and how accurate are the different sources of information they deal with and how to deal with uncertainties in the information they receive.

The tables below summarize the promotion requirements cited earlier and their relationship to the knowledge that we chose to elicit from the SMEs. The two tables address the two evaluation criteria of task performance level and task knowledge level.

Promotion Requirement	Knowledge Elicited
can do all parts of task	procedures, equipment used, forms used

Table 1: Task Performance Level

Promotion Req.	Knowledge Elicited
why task performed	relevant diseases
when task performed	triggering conditions
why steps needed	relationship of information accuracy to disease inference, how to handle uncertainties to maximize patient care, implications of deviations to expected intake/output based on patient condition

Table 2: Task Knowledge Level

The task performance criteria are more self-explanatory. In order to perform the task, it is important to know the procedures to use, the equipment to work with and how to record the information. The task knowledge requirement is more complex since it involves understanding the rationale behind the task. Our analysis of this requirement (which influenced the topics covered in the knowledge engineering sessions) is as follows. A patient has a certain medical condition which can be affected by the amount of fluid s/he takes in and puts out. This reasoning drove the portions of the elicitation that dealt with different types of disease. The condition influences both how often s/he should be monitored as well as signs that there is a problem with fluid intake and output. This drove the portion of the knowledge engineering that dealt with determining when the tasks should be performed and what were the triggering conditions that notified the technician that there was a need for monitoring. The information that the technician gathered would then be evaluated for implications regarding the patient's health. In performing the steps to gather information regarding intake and output, there is variability in how accurate this information is. The technician needs to understand this and make sound judgments in dealing with the information so as to make good inferences regarding implications for patient health. These considerations drove the portion of the knowledge elicitation that addressed the "why" the steps were important.

The above knowledge drives the assessment process. It is a blueprint for how to evaluate the technician's responses to the scenario presented. Each of the types of knowledge have implications for how to assess the technician. For example, if the technician omits checking the IV or the water pitcher, s/he cannot be given full credit for following the entire procedure. Similarly, if the technician records all information dutifully, but does not seek further diagnostic cues if there is a deviation in expected fluid intake, then s/he cannot be given full credit for understanding why the task is performed (i.e., its relationship to the larger issue of patient care).

As part of the knowledge engineering sessions with our SMEs we developed a scoring system. We made comments in the preceding paragraph about "full credit". Part of an assessment process is to come up with a scoring system that differentially scores varying levels of performance. This is an important requirement to insure that our technology can be adopted by an end user community. Specifically, the end user must be able to set the tasks to be assessed and the scoring criteria. The technology must adapt to the institution, not the other way around. Therefore, it was important to allow the Air Force to develop the scoring rubric and then RDC to implement it in our assessment technology.

The Measurement Of Performance Framework

This task is the glue between the knowledge engineering and the assessment system design. The knowledge engineering sets the requirement for what must be assessed. Development of the performance measurement framework creates the specifications for how the system will assess the required knowledge. Therefore, the framework must address the following issues:

1. How will the expert knowledge be represented in the system?
2. How will the system elicit the knowledge from the medical technician so it can be compared to the expert model?
3. How will this comparison be made and scored?

We focused on three primary requirements to developing a performance measurement system: 1) generating problems and scenarios that are complete and valid in their appropriateness for the skills to be assessed; 2) presenting the scenarios in a way that sustain the integrity of the scenario as people respond to scenario events; and 3) assessing performance based on the tested person's input.

In order to accomplish these requirements, three component technologies are needed. These are: 1) a mature framework for representing the domain and tested person's knowledge and performance that support evaluations against the performance criteria developed by the Air Force; 2) simulation technology that utilizes the domain model identified in 1) that allows for the scenario to unfold realistically based on the person's input; and 3) an assessment methodology that processes people's input, and evaluates it against performance criteria, and follows up with additional probes as needed to disambiguate uncertain evidence regarding level of performance.

We argue that the heart of this assessment tool is the expert model of the domain. It is used to define the requirements for what knowledge and skills need to be tested and therefore what scenarios and additional probes are needed. It is used to drive the simulation engine to determine how the scenario will unfold. Finally, it is used to evaluate the performance of those being assessed. We discuss each of these component technologies in turn.

Modeling Domain/expert And Trainee Knowledge. In the cognitive science and psychology literatures, several frameworks have been proposed as models of expert (and non-expert) knowledge. These schemes tend to address different types of knowledge. For example, scripts (Schank and Abelson, 1977) are used to represent goal and planning knowledge that is used in fairly routinized environments. Scripts are generalized sequences of steps used to achieve a goal.

Knowledge about data patterns and how objects are organized together can be represented by object frames (Minsky, 1975). Frames can be distinguished from semantic nets (cf., Quillian, 1966) which tend to organize information about individual concepts and relationships between them rather than collections of objects. For example, a hospital room may best be represented by a frame since it is a collection of people and equipment while a foley catheter may best be represented by a semantic net that describes its features.

Knowledge about situation-specific procedures can be represented by production rules (cf. Newell and Simon, 1972). Production rules are expressed in the form "IF [antecedent], THEN [consequent]", where antecedents are situational conditions that determine when procedures are to be executed and consequents are the procedures executed under those conditions. Production rules are useful in both carrying out procedures (e.g., "If this step has been completed, then do this next step.") and also generating inferences (e.g., "If the following problem features are observed, then infer that this is an [X] type of problem.").

Finally, causal and analogical reasoning can be captured by mental models (cf., Johnson-Laird, 1983; Leddo, Cardie and Abelson, 1987). In our framework, (Leddo, Cardie and Abelson, 1987), mental models are viewed as encoding the causal rationale for why a specific problem solving procedure is used. One of the factors that distinguishes the way experts solve problems from the way non-experts do is the former's heavy reliance on mental models and the ability to use them to select an appropriate problem solving strategy to meet a set of objectives.

We have discussed five different representation frameworks (scripts, object frames, semantic nets, production rules and mental models) for representing expert knowledge. Experts possess diverse knowledge that is richer that can be handled by any single framework (Leddo et al., 1990). Leddo, Cardie and Abelson (1987) developed an Integrated Knowledge Structure (INKS) framework that combines these individual schemes. In the INKS framework, scripts serve as the general organizer of knowledge, linking plans and goals together. Production rules give situation-specific procedures to be executed given conditions that arise during the execution of a plan. Frames organize collections of objects that are utilized in the execution of plans while semantic nets organize features of the individual objects within a frame. Mental models provide the rationale for why procedures are executed and how they are instrumental in achieving objectives.

The INKS framework is used to model medical service technician knowledge. For example, a script could represent the entire process of measuring and recording fluid intake and output. Production rules could model specific steps in the overall script such as select recording information from the IV display. Semantic knowledge could model knowledge about specific equipment. Frame knowledge could model knowledge about the hospital room itself and its layout. Mental models could model the rationale behind certain medical procedures are used and the causal relationship between diseases and symptoms.

In our project, a serendipitous event occurred. As noted earlier, the Air Force Career Field Education and Training Plan (CFETP) contains a scoring system for measuring performance. This scoring system lends itself well to the INKS knowledge components. For example, the task performance level requirements deal with progressively accurate procedural knowledge (from “Can do parts of the task” to “Can do the complete task quickly and accurately”). The most advanced level describes the entire task script.

The task knowledge levels correspond closely to components in the INKS framework. Level a is “Can name parts, tools and simple facts about the task”, which corresponds to semantic knowledge. Level b is “Can determine step-by-step procedures for doing the task”, which corresponds to production rules. Levels c and d refer to operating principles and advanced theory, respectively. These correspond to mental models.

Subject knowledge levels also correspond to the INKS framework. Level A is “Can identify basic facts and terms about the subject”, which again refers to semantic knowledge. Levels B, C, and D refer to principles, analysis, and evaluation, respectively, which are heavily mental model-oriented.

The Knowledge Elicitation Framework. Knowledge elicitation forms the heart of the assessment process (as well as be used to build the expert model). We believe performance assessment systems offer a unique opportunity for knowledge elicitation techniques in the role of training evaluation. The very nature of the simulation context is action-oriented so the system has ample opportunity to collect information on problem-solving behavior. Observing workers' behavior on the scenario allows us to see how they naturally solve problems as well as develop hypotheses about what they know and do not know. Question and answer-based knowledge elicitation techniques could be used to collect data to determine whether the hypothesized gaps in knowledge really do exist.

These two types of knowledge elicitation approaches are complementary. Observing problem solving has the strength of being used with realistic tasks. The elicitor poses little intrusion into the problem solving process. Hence, the resulting behaviors observed are likely to have high validity as a description of how the expert actually solves the problem. There are weaknesses with this approach, however. Simply observing an expert solve a problem does not necessarily shed light on his/her underlying cognitive problem solving approach. new problems.

Question and answer techniques attempt to address this deficiency by asking the expert pointed questions about how to solve a problem. The intention is to get the expert to explore his/her problem solving process and report it directly to the interviewer. A trained knowledge engineer has a distinct advantage in the question and answer approach compared to the protocol analysis approach. In the latter, the elicitor relies on the expert to show and explain all of the knowledge necessary to solve the problem. If the expert fails to do so, the elicitor may be left with serious gaps in the elicited knowledge. The question and answer technique allows the elicitor to assess what gaps s/he has in the elicited knowledge and asked specific questions to fill those gaps.

At Research Development Corporation, we have been working on a knowledge elicitation approach that is a hybrid of protocol analysis and question and answer techniques. The goal of this effort is to exploit the strengths of each approach while overcoming their shortcomings. Using protocol analysis allows the expert to demonstrate realistic, complete problem solving. Question and answer techniques allow the elicitor to fill in gaps in the observed process and seek explanations of the guiding principles behind the problem solving. Therefore, the hybrid technique has an expert solve a problem without interruption and then follows up the problem solving activities with questions based on the INKS framework to complete the knowledge model. Our knowledge elicitation approach is called Cognitive Structure Analysis (cf., Leddo and Cohen, 1989; Leddo, 1996).

Cognitive Structure Analysis (CSA) involves assessing people's problem solving goals, the strategies they use (what procedures, what sequence they occur in, etc.), the reasons behind these strategies, the features of the problem that are relevant, etc. CSA's probes are driven by both structural and content considerations of the knowledge structures in the INKS framework. CSA has been shown experimentally to be a powerful tool in modeling problem solving knowledge. In one study in which CSA was used to assess classroom knowledge, the assessments produced by CSA correlated .88 with students' performance in practical problem solving (Leddo and Sak, 1994).

Measuring Performance. In the section above, we discussed a knowledge elicitation approach to gathering information regarding performance and underlying knowledge. Once this knowledge is gathered, an overall assessment of performance is required. As discussed earlier, the scoring system needs to be defined by the user institution as they would be the ones that would want to make the determination as to what constitutes satisfactory performance. Performing this step requires that we map the knowledge model of the worker that the assessment tool generates to the scoring rubric provided by the Air Force so that a score can be generated.

In order to do this, we worked with the subject matter experts. They gave us criteria for reaching each level of performance cited in the CFETP. Fortunately, as discussed in the previous section, the Air Force promotion scoring system uses knowledge levels that are very similar to our INKS. Therefore, it was relatively easy to map the different factual, procedural and conceptual criteria described in the CFETP to our INKS framework so that a score could be generated. The benefit of our approach was that the score could be explained (and was in the assessment report produced for each worker) by referring to the INKS knowledge that was used to generate the score. This takes some of the “subjectivity” out of the scoring process that is often found in performance scoring systems.

An Automated Performance Assessment System.

We developed an automated performance assessment system that demonstrates each of the features and capabilities of our assessment system framework. The three primary capabilities we wanted to demonstrate were the use of the INKS framework as a means of assessing worker knowledge, the use of cognitive structure analysis (CSA) as a means of eliciting worker knowledge, and use of an organizationally defined scoring rubric as a means of providing the final rating. We supplemented this rating with a detailed evaluation of the worker, provided by the INKS model.

The system we developed was programmed in C++ and runs in Microsoft Windows 95. The system presents the user with scenarios in a simulated hospital environment, allows the user to perform the task of monitoring fluid intake and output, and then rates the user’s performance. The system uses both the user’s behaviors in the scenario as well as responses to CSA-driven questions to form its evaluation of the user. This evaluation of the user is then compared to the scoring rubric provided by the Medical Service Specialty CFETP (which was supplemented with discussions with SMEs to determine how to map user behaviors onto these scoring guidelines) to create a final rating.

We illustrate how this process works by providing a running example based on one of the actual scenarios we constructed. The user sits down at the computer and runs our assessment program. Once the program starts, the user selects a scenario from the scenario database. For illustration purposes, we will use as our running example, scenario 2, which is the scenario that assesses Task Knowledge (as defined in the Medical Service Specialty CFETP). When the scenario is loaded, two things happen. First, the current state of the world (i.e., the scenario parameters) are loaded into the world state database. This database keeps track of the current state of the scenario so that the assessment engine can evaluate user responses vis a vis the context in which they occur. Second, the initial scenario screen is presented to the user.

Here on the initial screen, the user is given the patient history. Since Task Knowledge relates to the causes behind the activities performed, the user must understand how the patient's medical condition may be affected by fluid intake and output. In this case the patient is being treated for congestive heart failure (which requires a reduction of fluid in the body) and is being given lasix (a drug that causes the body to output fluid in the form of urine). The first decision the patient must make is what time to return to the patient's room. This is entered on this screen. Once a user enters a response or initiates any action on the simulator, the simulator creates a "trigger" to notify the INKS-based expert system that an event has occurred. This trigger translates the event into a form readable by the INKS model. When the expert system receives the trigger, it queries the world state database to determine the context in which this action occurs. In this case, the user enters when he will make his/her next monitoring of fluid intake/output. This is a user-initiated event. The system queries the world state database and learns that lasix was given and what time it was given. Based on these two pieces of information, the INKS reasoner uses the INKS knowledge model to determine whether than is an appropriate time to monitor the patient (lasix normally takes ½ to 1 hour to work). If the INKS reasoner determines that a mistake has been made, this evaluation is sent to the user model to be included in the final evaluation report.

On the next screen, the user is actually in the patient's room. Here the user must measure fluid intake and output. As can be seen, the simulation depicts a patient in a hospital room. He is attached to an IV and has a Foley catheter. There is a tray near the bed with a water pitcher and a glass. The user is to complete the FLUID INTAKE/OUTPUT DATA RECORD (AETC FORM 1293, JAN 94).

There are no instructions on how to proceed. The user simply performs whatever operations in what ever order s/he wants to. The user navigates around the scene by using a mouse and clicking on objects that she or he wants to examine more closely. For example, suppose the user examines the foley bag to see how much urine is output. This event causes the simulator to create a trigger for this event to be sent to the expert system. Once received, the world state database is queried to help determine whether this is an appropriate action.

Because the user has examined the foley bag, the system expects, for efficiency sake, that the user will record his/her observation and then empty the bag. This is one case where the system has concrete expectations as to the order of procedure. The act of recording the amount of fluid in the foley bag is another event which sets off a trigger to the expert system. The expert system queries the world state database to determine whether the measurement is accurate. Once the user completes the recording, s/he should empty the bag. This action updates the world state database and resets the value of fluid in the foley bag to 0.

At this point, the user should realize that urine output is less than what would be expected, given that the patient received lasix. Since a goal in treating congestive heart failure is to increase fluid output, the user should be looking for signs of edema (swelling) on the legs or a gurgling sound on the lungs. Since both of these symptoms are present, the user should notify the nurse.

When the user is finished with the scenario, s/he clicks the "finished" button. At this point, the knowledge elicitor queries the user to gain additional information about his/her knowledge. As discussed earlier, the queries are based on the knowledge structures that comprise the INKS framework. The types of queries are constrained by the knowledge needed in order to evaluate the student according to the Task Knowledge criteria established by the Air Force.

Once the system's evaluation of the user is complete, it compares its user model to the scoring rubric established by the Air Force and fleshed out with the Air Force SMEs. Each of the activities that the user was expected to perform and each of the queries s/he was presented with were designed to provide feedback four each of the four Task Knowledge Level performance categories. For example, level a refers to "nomenclature" and relates to tools, parts, and facts about the task. This corresponds to semantic knowledge. The queries where the user had to click on items in the scenario provide diagnostic information about this scoring level.

Level b refers to “procedures” or whether the user can carry out the task. The actual performance on the simulation provided feedback regarding this. Specifically, the system determined whether all sources of information where checked and properly measured and recorded.

Level c refers to “operating principles” or the “why” behind the task. This was assessed by direct query. It is more difficult to assess this from behavior as the user can follow a procedure s/he was trained in and not understand why he is doing it. We believe this is a gap in many performance assessment approaches that look at behavior only but do not query the test taker for their understanding of operating principles.

Level d refers to “advanced theory” which is defined as “can predict, isolate, and resolve problems about the task”. We assessed this by presenting a situation where the evidence, given the patient’s medical condition, suggests that he is in some danger because of too much fluid retention. The way the system would assess this is by seeing whether, once the user saw the limited fluid in the foley bag, s/he checked for signs of fluid retention such as edema and gurgling on the lungs, and whether once those signs were present, s/he reported the patient’s condition to the nurse.

The cumulative information for each of these levels is aggregated by the tool. The tool uses the user model and the Air Force scoring rubric. The Air Force specifies that to achieve a given performance level, one must master all the performance levels below it. For example, one cannot receive a score of “c” (mastery of operating principles) which is sufficient for promotion if s/he misses factual or procedural knowledge.

The final score is sent to the simulator for presentation to the user. The simulator presents this final score with an explanation of where it found deficiency in the user.

REFERENCES

Brna, P., Ohlsson, S., & Pain, H. (1993) (Eds.). Proceeding of Artificial Intelligent in Education ‘93. Charlottesville, VA: Association for the Advancement of Computing in Education.

Greer, J. (1995) (Ed.). Proceeding of Artificial Intelligent in Education ‘95. Charlottesville, VA: Association for the Advancement of Computing in Education.

Johnson-Laird, P.N. (1983). Mental models. Cambridge, MA: Harvard University Press.

Kraje, B., Woolf, B.P., Fray, R., and Ghosh, R. (1993)., The EPRI Intelligent Tutoring System. Proceedings of the 10th Annual International Simulators Conference, Society for Computer Simulation: Washington, D.C.

Leddo, J. (1996). A knowledge assessment tool for diagnosing critical student learning needs. In P. Carlson and F. Makedon (Eds.) Proceedings of ED-Media 96-World Conference on Educational Multimedia and Hypermedia. Charlottesville, VA: Association for the Advancement of Computing in Education.

Leddo, J. and Cohen, M.S. (1989). Cognitive structure analysis: A technique for eliciting the content and structure of expert knowledge. Proceedings of 1989 AI Systems in Government Conference. McLean, VA: The MITRE Corporation.

Leddo, J. and Sak, S. (1994). Knowledge Assessment: Diagnosing what students really know. Presented at Society for Technology and Teacher Education, March.

Leddo, J. and Kolodziej, J. (1997). Distributed Interactive Intelligent Tutoring Simulation. Proceedings of the Interservice/Industry Training Systems and Education Conference. Arlington, VA: National Training Systems Association.

Leddo, J.M., Cardie, C.T., and Abelson, R.P. (1987). An integrated framework for knowledge representation and acquisition (Technical Report 87-14). Falls Church, VA: Decision Science Consortium, Inc., September.

Leddo, J., Cohen, M.S., O'Connor, M.F., Bresnick, T.A., and Marvin, F.F. (1990). Integrated knowledge elicitation and representation framework (Technical Report 90-3). Reston, VA: Decision Science Consortium, Inc.

Minsky, M. (1975). A framework for representing knowledge. In P. Winston (Ed.), The Psychology of computer vision. NY: McGraw-Hill.

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

Quillian, M.R. (1966). Semantic memory. Camridge, MA: Bolt, Beranek and Newman.

Schank, R.C. & Abelson, R.P. (1977). Scripts, plans, goals and understanding. Hillsdale, NJ: Erlbaum.