Interactive Learning Environments
2003, Vol. 11, No. 3, pp. 233–262

1049-4820/03/1103-233$16.00

Learning from a Computer Tutor with Natural Language Capabilities

Joel Michael[1], Allen Rovick[1], Michael Glass[2], Yujian Zhou[3] and Martha Evens[3]

[1]Department of Molecular Biophysics & Physiology, Rush Medical College, Chicago, IL, USA, [2]Department of Mathematics and Computer Science, Valparaiso University, Valparaiso, IN, USA, and [3]Department of Computer Science, Illinois Institute of Technology, Chicago, IL, USA

ABSTRACT

CIRCSIM-Tutor is a computer tutor designed to carry out a natural language dialogue with a medical student. Its domain is the baroreceptor reflex, the part of the cardiovascular system that is responsible for maintaining a constant blood pressure. CIRCSIM-Tutor's interaction with students is modeled after the tutoring behavior of two experienced human tutors. The effectiveness of CIRCSIM-Tutor has been tested by 50 first-year medical students. Using a pre-test/post-test paradigm we have demonstrated that significant learning occurs during a 1 hr interaction with the program. Students were also surveyed and indicated considerable satisfaction with the program.

INTRODUCTION

One-on-one tutoring is known to result in learning outcomes that are significantly greater than those produced by usual classroom activities (Bloom, 1984; Cohen, Kulik, & Kulik, 1982; Graesser, Person, & Magliano, 1995). The explanation for this phenomenon is not clear although Graesser et al. (1995) have argued that it is the collaborative problem solving that arises out of the dialogue between tutor and student that results in the learning gains that occur. Thus, even novice (peer) tutors can be quite effective in promoting learning (Graesser et al., 1995). Chi, Leeuw, Chiu, and La Vancher (1994) provide evidence that students who produce self-explanations in natural

Address correspondence to: Martha Evens, Computer Science Department, Illinois Institute of Technology, 10 West 31st Street, Room 236, Chicago, IL 60616, USA. Tel.: +1-312-567-5153. E-mail: evens@iit.edu

language consistently perform better on problem-solving tasks and also show better retention. This may also contribute to the apparent effectiveness of natural language tutorial dialogue.

Our intelligent tutoring system, CIRCSIM-Tutor, was built to help students learn to solve problems involving the negative feedback system that controls blood pressure in the human body. The system asks the students for predictions, evaluates those predictions to form hypotheses about student misconceptions, and carries on a natural language dialogue with the students designed to cause them to examine their mental models and revise them. Our motivation was to construct a kind of dialogue that allows the system to diagnose student errors and forces the students to give explanations.

In this paper we describe a classroom experiment with CIRCSIM-Tutor. The study was intended to determine whether the computer tutor capable of generating a natural language dialogue that we implemented would, in fact, help first-year medical students learn about the control of blood pressure. This experiment involved 50 medical students taking Physiology at Rush Medical College in a regularly scheduled 2-hr laboratory in November, 1998. They took a pre-test, spent an hour working with the Circsim-Tutor program, then took a post-test, and filled out a survey questionnaire.

The paper begins with a brief description of a precursor program, CIRCSIM (Rovick & Michael, 1986), use of which was determined to lead to increased understanding of the physiology of blood pressure regulation. We describe the human tutoring sessions that we have analyzed, which have together formed the basis for the development of CIRCSIM-Tutor. Then we present some examples of the dialogue generated by the system along with explanations of what the system is doing. These examples have been taken from an experiment with 50 students in a laboratory setting in November, 1998. We explain the design of the experiment, the pre-test, the post-test, and the survey questionnaire. We present the results of this experiment and discuss them. Finally, we describe some related research using natural language dialogue in intelligent tutoring systems, and discuss some future research.

THE CIRCSIM PROJECT

Problem-Solving in Physiology

Learning a scientific discipline like physiology involves acquiring a great many ‘‘facts’’ (vocabulary terms, definitions, concepts, bits of data, etc.) and

also learning algorithms with which one can use these facts to solve problems of a variety of types (Michael & Rovick, 1999). One type of problem that is very common in physiology, and also common in the professions in which physiology is employed, is predicting the qualitative behavior of a system that is disturbed (Kuipers & Kassirer, 1984). Learning to carry out qualitative, causal reasoning is a daunting task for many students.

This task becomes more difficult when the system being considered incorporates negative feedback (''homeostasis'') as is the case for most physiological systems (Dawson-Saunders, Feltovich, Coulson, & Steward, 1990). The presence of feedback increases the number and the complexity of the interactions between system variables that must be considered. Furthermore, the algorithm for successfully making qualitative predictions becomes more complex.

The Learning Task

The baroreceptor reflex (Berne & Levy, 1993) is that part of the cardiovascular system responsible for maintaining a more or less constant blood pressure. It is a complex system with many parameters interacting with one another, in many instances in ways that are counterintuitive. It is an important piece of physiology for which the ability to make qualitative predictions about system behavior is essential.

The Program

To assist students in understanding this system, and in developing the ability to predict system responses, Rovick and Michael (1986) developed a computer-assisted instructional program called CIRCSIM. Students are given the description of some perturbation of the cardiovascular system and are asked to predict the qualitative changes (increase, decrease, or no change) that will occur in seven cardiovascular parameters during the three time periods of the response: the Direct Response (DR) to the disturbance, the Reflex Response (RR), and the new, final Steady State (SS). The system analyzes these predictions to find errors in the predictions about the parameters and in the relationships between parameters. Then it picks a remedial paragraph from among 245 stored files and presents that text on the screen. The student must make all 21 predictions before any corrective feedback is displayed. When CIRCSIM was presented with a prediction table formed by taking the first column in Figure 2 and adding correct predictions in the rest of the table, the system listed two important equations, asked the

student to alter the predictions to satisfy these equations, and then displayed the following paragraph:

> A large reduction in blood volume (in this case a loss of 40% of the normal blood volume) results in so large a fall in blood pressure that the perfusion of the coronary circulation is reduced. That results in decreased contractility and further compromises the system's ability to maintain blood pressure. In all other respects the cardiovascular response to this decrease in blood volume is identical to the response to a smaller hemorrhage.

The interface used by CIRCSIM, which is called a prediction table (Rovick & Michael, 1992), helps the students to organize their thinking about the system (parameters are listed in a sequence that suggests the sequence in which changes occur and in which predictions should be made), and also makes it easy to deliver appropriate canned text to correct student errors and their underlying misconceptions (Michael, 1998). We found the prediction table concept so useful that we carried it over to CIRCSIM-Tutor.

Learning Outcomes with CIRCSIM

Using a pre/post-test paradigm the learning outcomes of student use of CIRCSIM were determined (Rovick & Michael, 1992). Four groups were tested: (1) a control group that did not use CIRCSIM but was tested at the same time as the treatment groups, (2) a group that used CIRCSIM in a solo mode – one student to a computer with no interaction between students allowed, (3) a group in which students worked together in pairs at the computer, and (4) a group in which students worked together in pairs at the computer in the usual computer laboratory setting, with the pairs free to interact with each other and with a circulating laboratory instructor (AR or JM). The results of the pre/post-test comparison showed that use of the program in any mode resulted in a statistically significant increment in learning, that pairs learned more than those working solo, and that interaction with the laboratory instructor resulted in the greatest learning.

In the course of this experiment two new examinations were written. One half of the students took the first examination as a pre-test and the second as a post-test, while the other half of the students took the tests in the opposite order. Pre-test performance on these two examinations was shown to be the same. These same tests were used in the experiment reported here.

In spite of this encouraging outcome, it was clear that CIRCSIM had only a limited ability to assist students to overcome certain fundamental misconceptions that one-on-one tutorial interaction often successfully eliminated. This was thought to arise from the inability of CIRCSIM to actually interact with the students. The CIRCSIM-Tutor project was born out of our belief that a computer tutor able to engage the student in a tutorial dialogue would be even more effective than CIRCSIM in helping students learn about the baroreceptor reflex and qualitative causal reasoning (Evens et al., 2001).

Analysis of Human Tutoring

In developing CIRCSIM-Tutor we have deliberately sought to simulate the tutoring behavior of our two experienced human tutors (JM and AR). More than 50 one-on-one tutoring sessions have been conducted with first-year medical students solving one or more CIRCSIM-like problems. These sessions were captured for analysis using a PC-to-PC communications program, the Computer Dialogue System (Li, Seu, Evens, Michael, & Rovick, 1992). The resulting transcripts have been analyzed to yield information about the sublanguage used by students and tutors to discuss the baroreceptor reflex (Seu et al., 1991), the knowledge used by the tutor (Khuwaja, Evens, Rovick, & Michael, 1992; Zhang, Evens, Michael, & Rovick, 1990) and the tutoring rules employed (Hume, Michael, Rovick, & Evens, 1996). More recently we have carried out extensive markup and used machine-learning techniques to discover tutoring rules (Kim, Glass, Freedman, & Evens, 2000; Shah, Evens, Michael, & Rovick, 2002; Zhou et al., 1999a, 1999b). We carried over the prediction table concept from CIRCSIM, but otherwise developed a totally new tutoring protocol modeled on human tutoring.

CIRCSIM-TUTOR

CIRCSIM-Tutor conducts a natural language dialogue with students solving a problem about the baroreceptor reflex. Students first make qualitative predictions about the responses of the system. The tutor then interacts with the student to correct any prediction errors that were made. The tutoring protocol and the tactics used to remediate student errors are a simulation of the tutoring carried out by our experienced human tutors (JM and AR). In the next section we present some extended samples of the dialogue between CIRCSIM-Tutor

and several students. We then give a brief description of the architecture of CIRCSIM-Tutor.

SAMPLES OF DIALOGUE FROM CIRCSIM-TUTOR SESSIONS

The best way to help readers understand how the student and the program interact, we believe, is to provide some examples from actual interactions recorded in November, 1998. Student #4 began by choosing the procedure named ''Hemorrhage – Remove 1.0 Liter.'' The full text of this procedure description is displayed in the ''Procedure Description'' window (upper right) on the screen seen in Figure 1. The initial instructions can be seen in the ''Tutoring Window'' on the upper left. In the samples of dialogue presented

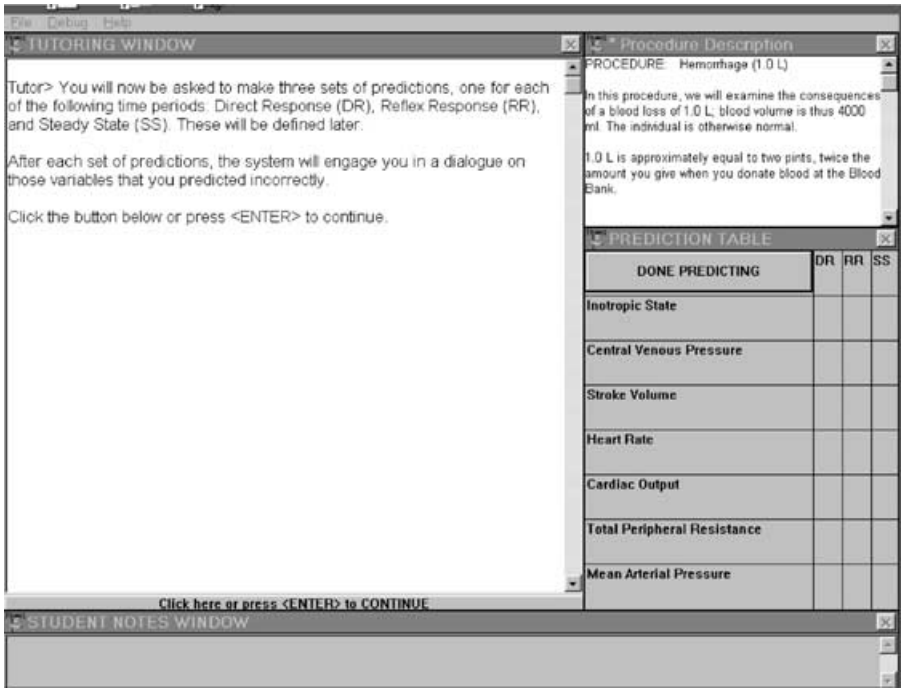

Fig. 1. A screen print from CIRCSIM-Tutor. The description of the procedure (the perturbation that is affecting the behavior of the cardiovascular system) can be seen in the window on the upper right.

here, the computer tutor output begins with a "**T**" and the student input with an "S". Comments and explanations to clarify what is occurring in the dialogue have been inserted. The tutor goes on to say:

T: The DIRECT RESPONSES occur before the baroreceptor reflex is activated. They produce the early changes in the affected variables. Enter predictions in the Prediction Table now. Make your predictions in a logical order, starting with the first variable affected. For each variable, click on the appropriate box and select a value from the menu: '+' represents an increase in value; '0', no change; and '−', a decrease. Please start making DR predictions.

The student now starts off with the prediction that Inotropic State (IS) will not change. Since we have observed that an error in predicting the response of the first parameter to change usually produces a large number of wrong answers (Cho, Michael, Rovick, & Evens, 1999), the tutor intervenes with a hint to make sure that the student begins the predictions with the correct variable.

This hint points to the physical location in the system where the first change occurs as a result of the perturbation. The goal of this hint is to help the student focus on the correct parameter. Unfortunately, in this instance, it does not succeed and the student's second try, a prediction that Heart Rate will increase, is also wrong. The tutor then gives a still more direct hint.

T: You have predicted IS. But IS is not the first variable in the table that is affected by the hemorrhage. When 1 liter of blood is lost, where in the circulation does the blood come from? Please try again.

T: No. HR is not the first one in the table that would be changed by this procedure. The hemorrhage decreases total blood volume and most of that decrease occurs in the venous compartment. Please try again.

The student now gets the point and enters the correct first prediction (Central Venous Pressure or CVP is decreased). The system then requests that the student enter the rest of the predictions for the Direct Response phase, analyzes the entries and marks the wrong answers with a slash across the Prediction Table entry. Figure 2 contains a snapshot of the screen as the tutor continues.

T: You will now be asked questions regarding the subject matter. Click on the Button or press <ENTER> when you have finished typing the answer.

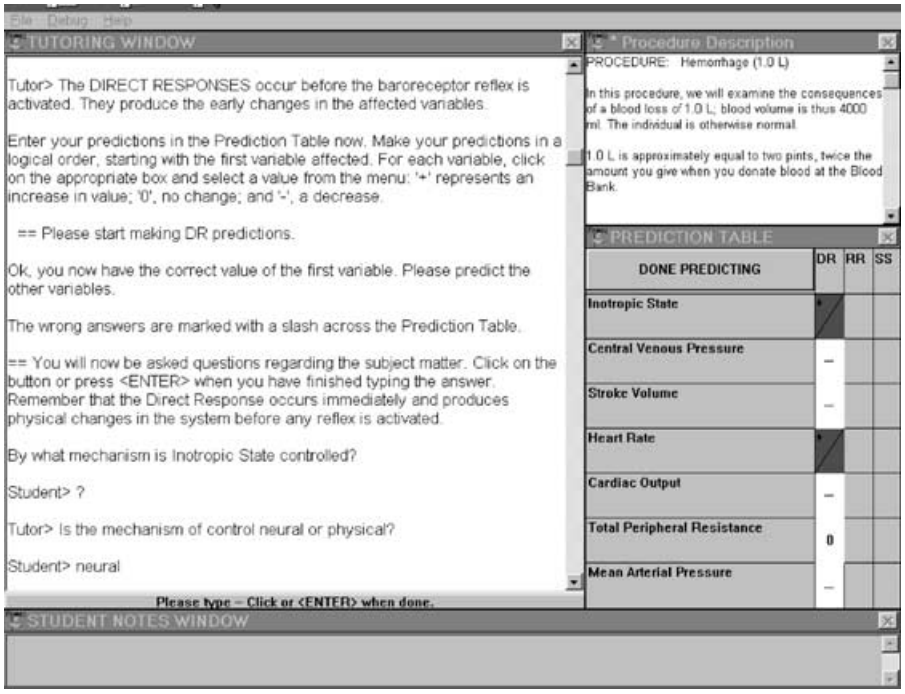

Fig. 2. A screen print showing the Prediction Table after the student has completed the DR column.

> Remember that the Direct Response occurs immediately and produces physical changes in the system before any reflex is activated.
>
> By what mechanism is Inotropic State controlled?

S: *?*

Early versions of CST crashed when a student answered a question with a question mark, but the tutor now responds by suggesting what kind of answer it is expecting.

T: Is the mechanism of control neural or physical?

S: *neural*

T: Correct, Inotropic State is controlled by the nervous system. What is the value of Inotropic State in DR?

S: *unchanged*

T: Correct, the value of Inotropic State is unchanged in DR. Please change the value of 'IS' in the predictions table.

The tutor now proceeds to deal with the other neural variables using a method designed to cause the student to infer the basic principle the student seems to be missing – ''Neural variables do not change in DR.''

T: Remember that IS, HR and TPR are all controlled by the nervous system. What is the correct value of Heart Rate?

S: unchanged

T: Correct, the value of Heart Rate is unchanged. Please change the value of 'HR' in the predictions table.

When the tutor has finished the remedial dialogue for the Direct Response phase, the wrong predictions are removed, and the tutor introduces the second phase of the response, the ''Reflex Response.''

T: Activation of the baroreceptor reflex is the key event in the negative feedback system that controls blood pressure. The Reflex Response (RR) represents the change in the values of the variables after the baroreceptor reflex fires. Your predictions in the RR column should represent the change in variables relative to their values in DR. Please start making RR predictions.

The student makes two erroneous predictions in this column, in Stroke Volume and Central Venous Pressure. Since Central Venous Pressure is a determinant of Stroke Volume, the tutor first deals with CVP.

T: You will now be asked questions regarding the subject matter. Click on the button or press <ENTER> when you have finished typing the answer. What is the determinant of Central Venous Pressure?

S: blood volume

This answer, ''blood volume,'' is an appropriate one in solving other problems, so the tutor agrees that this is a correct answer, but also explains that it is not relevant here.

T: Blood Volume determines CVP. However, no change in blood volume occurred since DR.

S: ok

This is the only situation where the tutor does not end the turn with a question. Clearly this is a mistake, because the student is confused about how to respond. The tutor is expecting to get a prediction table parameter as an answer and tells the student so.

T: Please respond with prediction table parameters

After the student figures out the right prediction for CVP, the tutor goes on to discuss Stroke Volume. It begins by asking:

T: What are the determinants of Stroke Volume?

S: central venous pressure and heart rate

The first answer is correct, but the second is not.

T: Well, Central Venous Pressure is one of the determinants, but Heart Rate is not.

The tutor now needs to generate a hint to help the student correct that error. The hint seen below is designed to get the student to think about the fact that all of the neurally controlled variables behave in the same way.

T: The other determinant, like Heart Rate, is neurally controlled. What is the other determinant of Stroke Volume?

S: inotropic state

T: Right, Inotropic State is one determinant of Stroke Volume. Stroke Volume is determined by Central Venous Pressure and Inotropic State. Which determinant is dominant in this case?

After a few more questions and answers the student completes correcting the predictions about the Reflex Response. The tutor now introduces the Steady State phase.

T: A new STEADY STATE is achieved when the system again stabilizes. Remember that though the system in the RR stage tries to compensate for the changes that happened in the DR stage, it never fully compensates. Your SS predictions should indicate the value change in the variables relative to their value before the onset of the original perturbation.

This particular student correctly predicted all the SS responses. The tutor acknowledges this and goes on to summarize the responses of the baroreceptor

reflex to the disturbance that was described in the problem. In the current version of CIRCSIM-Tutor this summary is stored.

T: OK, now that you have completed this procedure let's summarize the physiological mechanisms that are involved with the regulation of blood pressure. The decreased blood volume directly reduces central venous pressure. This reduces filling pressure and end diastolic volume. Thus, stroke volume falls, decreasing cardiac output. This in turn causes mean arterial pressure to fall. Activation of the baroreceptor reflex follows the fall in blood pressure and increases heart rate and Inotropic State. Arterial resistance is also increased and thus mean arterial pressure is restored towards its normal level. Additional compensatory mechanisms involving (1) the absorption of interstitial fluid due to a fall in capillary pressure and (2) ADH and renin-angiotension-aldosterone acting via the kidneys eventually restore blood volume and hence cardiovascular function. However, these changes are slow and do not occur within the time period of this experiment. Thus, loss of small volumes of blood produce minimal, and relatively short-term, effects. End of the SS tutoring session. Type return to continue.

We added these long summaries because students who had previously used the CIRCSIM program asked for them. Students in our present test population seem satisfied with the output seen here, but we are concerned because we are not convinced that reading a long paragraph is the best use of student time. Human tutors sometimes briefly summarize problems that the student has surmounted. More often, they construct much more interactive summaries as a series of short questions to the student using the strategy that Sanders (1995) named a ''Directed Line of Reasoning'' also discussed in (Hume et al., 1996).

Another student (#5) got all the DR predictions right but made the same error on SV in RR as the previous student. We show the remediation dialogue for this variable in order to show how the system copes with spelling errors, strange syntax, and symbols substituted for words. Here the tutor starts out the same manner seen in the example above, but the student mistypes ''Inotropic state'' and just runs the two names together without using a punctuation mark or a conjunction. The system fixes the error without commenting on the spelling or the syntax and constructs a typical response.

T: What are the determinants of Stroke Volume?

S: *notropic state central venous pressure*

T: Correct, Stroke Volume is determined by Inotropic state and Central Venous Pressure. Which determinant is dominant in this case?

The dialogue continues and the student starts to use the plus and minus symbols from the Prediction Table in the dialogue in place of ''increase'' and ''decrease.'' We expect this and have added these and many other symbols to the lexicon.

S: *central venous pressure*

T: Correct, the important determinant of Stroke Volume in this case is Central Venous Pressure. Is the relationship from Central Venous Pressure to Stroke Volume direct or is it inverse?

S: *direct*

T: Good, you got the correct answer. What is the correct value of Stroke Volume?

S: –

T: Correct, the value of Stroke Volume is decreased. Please change the value of 'SV' in the predictions table.

Any system that tries to hold a dialogue with a human being must expect terse answers. The work of Thompson (1980), among others, led us to expect this behavior and design for it from the beginning. The system is able to recognize the last student answer ''−'' as meaning ''Stroke Volume will decrease'' because it is expecting an answer to the question it just asked ''What is the correct value of Stroke Volume?'' The possible answers are ''increase'' or ''decrease'' or ''no change.'' The system will understand ''D'' or ''Decrease'' or ''Down'' or ''minus'' or ''−'' as equivalent to the complete sentence and produce the same logic form.

In addition, students also integrate pieces of equations into sentences. We observed many instances of this integration of algebra and English in human tutoring sessions (Glass, 1999, 2001), and decided we might as well go ahead and generate such language ourselves to signal to students that it is appropriate to use it with the system as in the following excerpt from the dialogue between the system and Student #10:

T: What are the determinants of Cardiac Output?

S: *hr*

T: Right, Heart Rate is one determinant of Cardiac Output.
CO = HR x what?
What is the other determinant of Cardiac Output?

S: *sv*

T: Right, ...

Sometimes students come up with a correct answer but one that is not what the tutor was looking for. We classify this kind of answer as a ''near miss.'' We do not want to tell the student that the answer is wrong since this is not the case. Instead we want to elicit the expected answer. Student #10 produced the typical ''near miss'' response below. The system asked a question designed to elicit the expected answer.

T: What is the determinant of Central Venous Pressure?

S: *central blood volume*

T: And what parameter in the prediction table determines Central Blood Volume?

S: *cardiac output*

T: Right, CO determines CBV, therefore determines CVP.

This student got the point immediately. But another very similar interchange found the student unable to figure out the right response. CIRCSIM-Tutor therefore goes on to give the student the answer since it has already hinted twice.

One of the most serious misconceptions that students reveal in this area is a confusion between Inotropic State and the Frank-Starling effect. We are particularly proud of the way our system diagnosed this misconception, explained the problem, and produced a successful hint, in the example below. This approach is still much less interactive than the one taken by experienced human tutors, however. In the next version of the system, we want to make CIRCSIM-Tutor give fewer long explanations and attempt to elicit more information from the student.

T: By what mechanism is Inotropic State controlled?

S: *ventrical filling*

T: No, you are confusing the Frank-Starling effect (increased filling gives greater output) and contractility. Inotropic State (Contractility) is the

force of contraction at any given filling or at any end diastolic volume. An increase in sympathetic stimulation of the myocardium increases the Inotropic State of the muscle. So, it is neurally controlled. Remember the reflex will not activate until the RR stage. What is the value of Inotropic State in DR?

S: 0

T: Correct, the value of Inotropic State is unchanged in DR.

Currently the system is available for casual student use in the computer laboratory at Rush. We have added three more procedures and we are trying out a number of types of hints and other responses to unexpected answers. We are hoping to discover more problems with discourse or instructional planning before we carry out another large experiment.

ORGANIZATION OF THE CIRCSIM-TUTOR SYSTEM

CIRCSIM-Tutor has an Expert Problem Solver and Domain Knowledge base, a Student Modeler, and a Pedagogical Module, like any other Intelligent Tutoring System (Anderson, 1988; Anderson, Corbett, Fincham, Hoffman, & Pelletier, 1992; Burns & Capps; 1988). Because of the need to carry on a natural language dialogue, the interface has been split into three components: the Input Understander, the Text Generator, and the Screen Manager. The need for dialogue planning as well as lesson planning convinced us to replace the pedagogical module with a fullblown hierarchical planner.

As the remedial dialogue begins, the Instructional Planner (Woo, 1992) calls the other components in turn, beginning with the Screen Manager. The Screen Manager displays messages to the student (Brandle, 1998), collects any text that the student types in, and returns it to the Instructional Planner. The Instructional Planner sends the input text to the Input Understander, which returns a logic form representation. This logic form is sent to the Student Modeler, which checks whether the answer is correct or not and updates the Student Model (Zhou & Evens, 1999). The Instructional Planner plans the response and calls the Text Generator to turn it into words. The Text Generator puts the response together a sentence at a time and returns it to the Instructional Planner, which calls the Screen Manager to output the turn to the student and waits for student input, as the cycle begins again.

Given the predictions input by the student in Figure 2, the Instructional Planner sets up as a lesson goal to tutor the student on the relationship between Central Venous Pressure and Stroke Volume. One rule for tutoring relationships sets up four subgoals: to establish the determinants of SV, decide which determinant is dominant in this situation, discuss whether the relationship is direct or inverse, and then determine the value of Stroke Volume. The Planner then calls the Text Generator to turn this plan into text one sentence at a time. The dialogue is definitely limited by the fact that it is planned a sentence at a time, while human tutors clearly plan much farther ahead.

The Input Understander parses the natural language input from the student, using cascaded finite state automata, produces a logic form representation of its meaning, and returns it to the Planner (Glass, 1999, 2000, 2001) calling for spelling correction as needed (Elmi & Evens, 1998). The input understander recognizes a number of different types of student answers, including partially correct answers, near-misses, ''grain of truth answers,'' and answers revealing some kinds of misconceptions. The dialogue above showed the plan generated when a student gets an answer partially correct; here, the machine tutor acknowledges the correct part of the answer and gives a hint to help the student figure out the other part. Suppose the student had answered ''preload'' instead. An earlier version of the system told the student that this answer is wrong. But this answer is not wrong. While it is not the prediction table parameter that the system expects, it is on the correct solution path. It is what we call a ''near-miss'' answer. Now the machine tutor responds to the answer ''preload'' with:

T: Right, preload determines Stroke Volume.
What parameter in the prediction table represents preload?

Another important category of answer is the ''Grain of Truth'' answer (Woolf, 1984). Here the tutor focuses on whatever is relevant in the student's answer and leverages that to get an answer to the original question. If the student answers the question about the determinants of Stroke Volume with ''Heart Rate and Cardiac Output,'' the system responds with:

T: Well, it is true that CO = SV * HR. But what I was asking is what determines how much blood is ejected from the heart each time it beats (the SV)?

These response plans are generated from rules on the fly, but the system has some prestored plans to deal with serious misconceptions. If the student gives the answer ''I don't know,'' the system responds with:

T: Remember, the central venous compartment is very compliant. Again, what determines Stroke Volume?

This same hint along with a negative acknowledgement is produced if the student gives an entirely erroneous answer.

The Text Generator receives logic forms from the Planner and turns them into sentences. The generator was written by Ru-Charn Chang (Chang, Evens, Michael, & Rovick, 1994) and uses a Lexical-Functional Grammar approach. Yujian Zhou et al. (1999a, 1999b) developed the rules for giving hints and answers to new question types recognized by the new Input Understander.

Circsim-Tutor was originally written in Procyon Common Lisp on Macintosh computers. Now that both the U.S. Navy and Rush Medical College are using Wintel-based personal computers we have ported the program to PC's running Allegro Common Lisp. A copy of the program can be obtained by sending email to Martha Evens.

STUDENT USE OF CIRCSIM-TUTOR

The Subjects

Fifty first-year medical students enrolled in the Medical Physiology course participated as paid volunteers in the experiment. They had already completed all scheduled class sessions on cardiovascular physiology (lectures, a computer laboratory on hemodynamics, and two problem solving workshops) and were scheduled to do a computer laboratory on blood pressure regulation the next day.

The Protocol

Half the students worked in pairs at a computer and half worked alone at a computer. All students were given 30 min to complete a pre-test (see below). They then worked for 1 hr with CIRCSIM-Tutor, completing as many of the four available procedures as possible in the allotted time. Forty-seven students completed all four of the available procedures, and three students completed three procedures. All students then completed the post-test (see below) and a survey (see below) in the remaining 30 min.

The Pre- and Post-Tests

The pre- and post-tests were each composed of three parts. They had been validated in a previous experiment (Rovick & Michael, 1992).

In Part 1 the students were asked to describe the relationships that are present between eight cardiovascular variables (''variable 1 directly/inversely determines the value of variable 2''). This represents an assessment of the students' knowledge about the component qualitative causal processes that make up the baroreceptor reflex. There are a total of 12 relationships and the maximum number of points on this component of the test is 24 (1 point is awarded for correctly indicating the existence of a causal relationship between two variables and 1 point is awarded for correctly identifying the nature of that relationship). Part 1 was identical in the pre-test and post-test.

Part 2 is a Prediction Table to be filled in describing the responses of the system to a described perturbation (a malfunctioning pacemaker increases an individual's heart rate from 70/min to 120/min, or decreases heart rate from 70/min to 50/min). These two problems have been used before as pre- and post-tests and are equally difficult to solve. Such problems test the students' ability to apply their knowledge of the component qualitative causal processes (relationships) to predict the response of the system to a disturbance. There are 21 possible points (correct predictions) for this component of the test. We also identified seven different categories of misconceptions (some having multiple instances in any one problem) that can occur in students' responses (Rovick & Michael, 1992) and counted the number of categories represented and the total number of misconceptions present in each student's solution.

Part 3 of the test consists of four multiple choice questions describing clinical situations that involve the same relationships between variables that are a part of the CIRCSIM-Tutor problems; there are four points for this part of the test. This component of the tests assesses the students' ability to apply their understanding of the responses of the system in a different, more clinically based context. The multiple choice questions used in the pre- and post-tests are of comparable difficulty.

The Survey

The survey (see the Appendix) contained 10 statements intended to probe students' attitudes about various aspects of the program with a 5-point Likert scale ranging from 1 = ''definitely YES'' to 5 = ''definitely NO.'' Space was provided for written comments. We were seeking feedback from the students about the ease of using the system and its interface, as well as comments on the perceived utility of the program as a learning resource.

RESULTS

Robustness of the Program

During this experiment the 50 students generated 1801 turns. The spelling correction program corrected 30 spelling errors including ''cariac'' to ''cardiac,'' ''trp'' to ''tpr,'' and ''inotrphic'' to ''inotropic.'' The system failed to understand 29 turns, but in every one of these it explained to the student what kind of answer it was expecting (as in the example above: ''Please give a prediction table parameter as a response.'') and the student was able to produce a reasonable answer. When we examined these 29 turns, we could not understand what the student meant in 10 of these 29, but we analyzed the other 19 failures with the goal of devising better responses. The system failed to recognize ''neurological'' and ''central venous volume'' and we have added these to the lexicon. It failed to break ''istpr'' down into ''is'' followed by ''tpr'' because it does not handle joins well. It also failed to recognize ''metabolic factors,'' which is really outside the domain of the system. Four of the 19 failures involved expressions of frustration (''kiss my ass'' was the mildest). The system responded to these also with a statement about the expected input but we would like to figure out a more effective response strategy. The simplest might be to give the student the correct answer at this point and go on to the next item on the agenda.

Learning Outcomes

Analysis of the pre- and post-test results clearly demonstrate that use of CIRCSIM-Tutor for 1 hr results in learning about the baroreceptor reflex and a greater ability to use this understanding to carry out qualitative reasoning about the system (see Table 1). Students were able to correctly describe more of the relationships between system variables (Part 1), with scores increasing from 13.64 on the pre-test to 16.16 on the post-test (t test, $p < .001$). The students were also better at making qualitative predictions about the responses of the baroreceptor reflex (Part 2). A comparison of pre- and post-test predictions showed many more corrections of errors (wrong on pre-test but correct on post-test) than the reverse (correct on the pre-test but wrong on the post-test) with this difference being statistically significant ($p < .001$ using the McNemar change test; Siegel & Castellan, 1988). Students also decreased the number of misconception categories present in their predictions from 2.67 to 1.28 (t test, $p < .001$) and decreased the total number of misconceptions from 4.07 to 1.83 (t test, $p < .001$). Finally, they were also

Table 1. Use of CIRCSIM-Tutor Improves Student Performance on All Three of the Assigned Tasks.

Test component	Pre	Post	Significance
Relationship Points (Part 1)	13.64	16.16	$p < .001$ (t test)
Predictions (Part 2)			
Wrong-correct	197		$p < .001$ (McNemar)
Correct-wrong	67		
Misconception categories	2.67	1.28	$p < .001$ (t test)
Total misconceptions	4.07	1.83	$p < .001$ (t test)
Multiple choice questions (Part 3)	2.24	2.96	$p < .001$ (t test)

better able to use their understanding of the baroreceptor reflex to answer clinically-based multiple choice questions (Part 3); scores went from 2.24 on the pre-test to 2.96 on the post-test ($p < .001$).

Survey Results

Student responses on the survey were generally quite positive about their experience with CIRCSIM-Tutor. They believe that the program helped them understand the baroreceptor reflex and that it helped them learn to predict system responses (see Fig. 3). They were less positive about the quality of the general dialogue and also the quality of the hints and explanations, but here

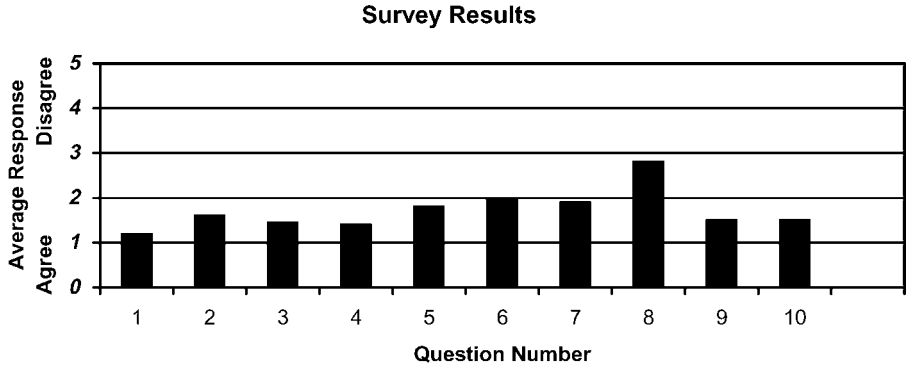

Fig. 3. Distribution of the students' responses to the survey following their use of CIRCSIM-Tutor for 1 hour. The questions can be seen in the Appendix.

too, their opinion was favorable. They were quite divided in their opinion of the protocol used by CIRCSIM-Tutor (requiring that they complete the predictions of an entire column before receiving any feedback); some students liked it and others did not.

Table 2. Student Comments from the Survey, Both Positive and Negative (33 Students Wrote Comments, and 8 of them Dealt Largely with the Physiology Contents of the Program).

There were a few instances when the dialogue appeared contradictory, yet one couldn't ask questions or further explore it, so in those cases I remained unclear on the particular situation. Otherwise very helpful program.
It would be helpful if the system provided a little more explanation about my wrong answers. I felt I corrected my errors because it told me to, but I didn't always understand the explanation. For example, I made a mistake in the SS segment.
The system told me that SS followed RR and therefore the answer would be similar to DR.
The fact that one could use abbreviations was pretty ????? and helpful; specific word format was not an issue. The explanation for steady state answers and how it followed the direct response or baroreceptor response was a little weak. Otherwise, excellent program!
I enjoyed using this program – it was fun and it helped clarify certain things for me.
I thought the program was excellent – especially the way it responded to individual responses precisely. I think it would be very helpful.
It may be only my personal view, but whenever information, particularly when there is a lot of it, appearing on a computer screen, I tend *not* to read any of it. If you use simple short phrases for responses I might read them with more attention.
A deeper explanation of why answers are incorrect should be given. The print of the scenario described was too small. Overall it was an easy to follow program with concise answers.
Somewhat hard to answer questions presented because I didn't understand what was being asked.
The questions asked by the tutor are not phrased clearly and acceptable answers should be displayed along with the question. Too many questions not enough explanation.
Very good – I would like to have of these situations (more than the 4 we covered). Please make this available before the exam. This is the best way to understand cardiovascular physiology.
A flow chart would have been helpful.
The explanations could be a little more detail & informative. Otherwise, great program.
BRAVO! Will it [the program] be available before the final exam?

The students' written comments (Table 2) were often quite enthusiastic about the program and its value as a learning resource. Such comments obviously parallel their numerical ratings. Their major criticism was leveled at the quality of the explanations that CIRCSIM-Tutor delivers when critiquing wrong answers. The program cannot yet describe the physiology underlying the causal relationships being tutored.

DISCUSSION

The experiment just described was designed to discover whether students would, in fact, learn from a computer tutor that uses natural language to conduct a dialog about the mechanisms responsible for blood pressure regulation. We believe that tutoring is best mediated by natural language. The students need to learn the language of physiology as they learn its concepts. The tutor can tell what the students do and do not understand much better from their own choice of language. Most importantly, students learn more if they express these new ideas in their own language (Chi et al., 1994). Two recent experiments have shown that motivating students to produce self-explanations results in improved learning outcomes (Aleven, Koedinger, & Cross 1999; Conati & VanLehn, 1999).

Chi et al. (2002) classify tutoring episodes on two scales: interactivity and constructiveness. An episode is interactive if the student is forced to respond. It is constructive if the student is led to construct knowledge actively. Our human tutoring sessions are both highly interactive and highly constructive. The tutor's goal is to help the students learn to solve problems. The tutors do not provide answers; they ask questions and give hints. They rarely give explanations; instead, they elicit explanations from the student. We have tried to make the CIRCSIM-Tutor system as interactive and constructive as possible given the limitations of machine understanding. These limitations have forced us to avoid asking questions that are likely to produce long open-ended answers that we cannot parse, but otherwise we have tried to produce dialogue as like that of the expert tutors as possible.

The idea that tutoring involving a natural language dialogue between tutor and student will lead to greater learning is supported, although not yet proven, by the results reported here. Students did learn from CIRCSIM-Tutor. The students also report that they like using the program because they feel that they are learning from it. They not only do better in solving the same kind of

problem in the same domain, they also transfer this ability to use causal reasoning to more explicitly clinical problems presented in a quite different format.

Perhaps we should explain the assumption of independence that underlies the statistical analysis here. Some students (approximately half) worked in pairs because there were not enough computers available for all students to work alone. Both in this study and in the 1992 study of CIRCSIM, we assumed that the students who worked with the system in pairs were independent. They certainly took the pre-test and the post-test independently. In spite of the extra work that was entailed in the test-scoring process, we gave the students these tests in paper form and proctored them to ensure that they solved the pre-test and post-test problems separately. Thus, we assume that the results reflect each individual's level of competence at the time of testing. Readers might also question the absence of a control group ''doing nothing'' (i.e., learning physiology on their own). We did use a control group ''doing nothing'' in the CIRCSIM study (Rovick & Michael, 1992) and a control group reading a specially prepared text in an unpublished study of the effectiveness of human tutoring. In both cases the treatment groups did significantly better than the control groups.

One major objective in moving from CIRCSIM to CIRCSIM-Tutor was to incorporate features that would support dialogue generation, but there are many other differences between the two programs. To reach our goal of generating a tutorial dialogue we added a dynamic student modeler, a problem-solver, and an instructional planner that does adaptive dynamic planning of the tutoring session. CIRCSIM-Tutor uses the student model to adapt the remedial dialogue to the student's needs and it changes its plans, or drops old plans and makes new plans, on the basis of the student's predictions and responses in the dialogue. These features allow CIRCSIM-Tutor to carry on a tutorial dialogue using natural language, with the tutor able to understand the student input and generate appropriate responses. The tutor now provides a variety of different hints in response to different types of student errors. It can distinguish between ''near miss'' answers where a follow-up question may elicit the right answer, partially correct answers, and ''grain of truth'' answers. It can make sense of answers in which algebraic symbols are included in English sentences. When it does not understand the student it provides feedback indicating what kind of input is expected.

CIRCSIM-Tutor can now carry on natural language dialogues with students that lead the student through the problem-solving process and provide

favorable learning outcomes even though they are not as sophisticated as dialogues generated by human tutors. We have shown that it is feasible to build a tutor that interacts with the student in natural language. We have developed a methodology for the analysis of human tutorial dialogues and the development of tutoring rules to synthesize similar dialogues (Kim et al., in preparation).

OTHER TUTORING SYSTEMS USING NATURAL LANGUAGE DIALOGUE

The power of natural language as a medium for tutoring has been recognized for many years, and some of the earliest attempts to apply the techniques and concepts of artificial intelligence to teaching were language based. The first intelligent tutoring system to carry on a natural language dialogue with the student was SCHOLAR (Carbonell, 1970). SCHOLAR was able to answer questions from the student involving the geography of South America and it also had a list of topics to raise with the student. It made only limited tutorial plans and no dialogue plans, but if the student interrupted a question by asking a question SCHOLAR would raise the issue again later. After Carbonell's untimely death, Collins and Stevens continued their joint work on dialogue-based systems and produced a more sophisticated tutor named WHY (Collins, Warnock, & Passafiume, 1975; Stevens & Collins, 1980).

WHY deals with causal reasoning in the domain of meteorology, and its focus was also on helping the student learn to solve problems, so its approach is very relevant to our task. In an early version of WHY, knowledge of the domain was represented in a hierarchy of scripts to capture stereotypical sequences of events (Stevens & Collins, 1980). Scripts were used not only to generate responses but also to understand the student's mental models and misconceptions from the student's language (Stevens, Collins, & Goldin, 1982).

Another important series of experiments with natural language in tutoring systems was carried out by Brown, Burton, and deKleer (1982) at Xerox PARC. The Sophie system provided qualitative explanations of the meaningful measurements and decisions that must be made in the troubleshooting of electronic networks, but computer limitations reduced the extent of natural language processing it could handle.

Graesser and his students at the University of Memphis have taken a totally different approach to generating a tutorial dialogue via AUTOTUTOR, which

tutors students in computer literacy (Person, Bautista, Graesser, Mathews, & Tutoring Research Group, 2001). Instead of parsing the input it uses Latent Semantic Analysis to compare the student's answer with a set of ideal answers. It generates a number of different kinds of dialogue moves including prompts, hints, corrections, and summaries, but its focus is on asking questions that force the student to construct explanations.

A more recent conversational system is CREANIMATE (Edelson, 1996), developed at the Institute for the Learning Sciences. CREANIMATE is designed to teach animal adaptation: the relation between the animal's morphology and what it can do. The system asks the student to propose modifications to animals. For example, a student wants to create a monkey with wings. CREANIMATE offers to show the student a fox bat (a mammal with wings). Then it asks a question:

T: If your monkey is going to have wings, that should help it do something. Why would you like your monkeys to have wings? So it can ....

S: fly away from its enemies.

Cawsey's EDGE system (1992) tutors the student about electrical circuits. It generates interactive explanations of the kind that we are trying to produce. It also generates follow-up questions and responses to questions from the user. The dialogue planning in EDGE is very sophisticated, but the actual sentence generation, like that in CREANIMATE and in Woolf's (1984) MENO-Tutor, is all template driven; none of these programs uses modern generation techniques.

Sophisticated generation can be found in the work of Moore and her students. The PEA system (Moore, 1995) is designed to help users improve their Lisp programs. The system asks what aspect of the program needs to be changed, then it makes recommendations and offers explanations. PEA maintains a user model and a dialogue history. Moore, Lemaire, and Rosenblum (1996) also carried out a series of experiments using the past history of the tutoring session to recall previous problem-solving attempts by the student. This required the generation of retrospective language and new kinds of tutoring schemas to generate the dialogue, as well.

Another very interesting tutoring system is still under construction at the University of Pittsburgh. ANDES (Freedman, Penstein Rosé, Ringenberg, & Van Lehn, 2000; Gertner, Conati, & VanLehn, 1998; VanLehn et al., 2000)

takes a very different approach to student modeling but it is also aimed at teaching causal reasoning using language. As far as we know, it is also the only other dialogue-based tutoring system to be tried on a significant number of students. The same group is now fielding a tutor with natural language capabilities based on Freedman's Atlas Planning Environment or APE (Freedman, 2000a, 2000b). We are now starting to build a new version of CIRCSIM-Tutor ourselves that uses APE as its planner.

FUTURE RESEARCH

We need to carry out a further study of CIRCSIM-Tutor that compares its effectiveness with that of CIRCSIM. We should also compare the students' performance after using the system with a control group that reads material from a relevant text. Since the system is being used as part of a course it might be possible to measure whether the students retain material better after using the system than they do after reading this material.

We have been working on a new version of CIRCSIM-Tutor (the third so far), because we want to improve the discourse planning in a number of ways. We want to add more ways to express all the current tutoring tactics. We need to plan the discourse over multiple turns, if we are ever to approach the verbal sophistication of our human tutors. Also, we need to make bigger and better discourse plans to be able to deliver the kind of explanations of underlying mechanisms that the students are requesting. We are working on further studies of discourse plans and tutoring language in human tutoring sessions (Kim et al., 2000). We are also working on a further comparison of the differences between novice and expert tutors, extending the work in (Glass, Kim, Evens, Michael, & Rovick, 1999).

We also want to be able to experiment with different tutoring protocols in order to answer some more fundamental educational questions. The protocol used in the current version of our system postpones commenting on most student predictions until a whole prediction table column is complete. We chose this protocol because it allows us to build a much richer student model. Some students have asked for immediate feedback and conventional wisdom would appear to support this notion. We would like to examine these alternatives in the framework of learning outcomes with CIRCSIM-Tutor because we can carry out a controlled experiment much more easily with a machine tutor than with human tutors.

There is another question we want to explore with a machine tutor: Do students learn problem-solving algorithms better if they are taught these algorithms explicitly (stated in direct language) or if they are left implicit as they are in the current version of the systems. Most of all, we want to carry out yet further experiments to demonstrate the value of natural language in machine tutoring.

ACKNOWLEDGMENTS

This work was partially supported by the Cognitive Science Program, Office of Naval Research under Grant 00014-00-1-0660 to Stanford University as well as Grants No. N00014-94-1-0338 and N00014-02-1-0442 to Illinois Institute of Technology. The content does not reflect the position of policy of the government and no official endorsement should be inferred. Conversations with Dr. Reva Freedman first at LRDC and now at Northern Illinois University have been of great assistance in our work.

REFERENCES

Aleven, V., Koedinger, K.R., & Cross, K. (1999). Tutoring answer explanation fosters learning with understanding. In S.P. Lajoie & M. Vivet (Eds.), *Proceedings of AIED-99* (pp. 199–206). Amsterdam: IOS Press.

Anderson, J. (1988). The expert module. In M.C. Polson & J.J. Richardson (Eds.), *Foundations of intelligent tutoring systems* (pp. 21–54). Hillsdale, NJ: Erlbaum.

Anderson, J.R., Corbett, A.T., Fincham, J.M., Hoffman, D., & Pelletier, R. (1992). General principles for an intelligent tutoring architecture. In J.W. Regian & V.J. Shute (Eds.), *Cognitive approaches to automated instruction* (pp. 81–106). Hillsdale, NJ: Erlbaum.

Berne, R.M., & Levy, M.N. (Eds.). (1993). *Physiology* (3rd ed). St. Louis: Mosby Year Book.

Bloom, B.S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-on-one tutoring. *Educational Researcher*, *13*, 4–16.

Brandle, S. (1998). *Using joint actions to explain acknowledgments in tutorial discourse: Application to intelligent tutoring systems*. Unpublished doctoral dissertation, Illinois Institute of Technology, Chicago.

Brown, J.S., Burton, R.R., & deKleer, J. (1982). Pedagogical, natural language, and knowledge engineering techniques in SOPHIE-I, II, and III. In D. Sleeman & J.S. Brown (Eds.), *Intelligent tutoring systems* (pp. 227–282). London: Academic Press.

Burns, H.L., & Capps, C.G. (1988). Foundations of intelligent tutoring systems: An introduction. In M.C. Polson & J.J. Richardson (Eds.), *Foundations of intelligent tutoring systems* (pp. 1–19). Hillsdale, NJ: Erlbaum.

Carbonell, J.R. (1970). AI in CAI: An artificial intelligence approach to computer-assisted instruction. *IEEE Transactions on Man-Machine Systems*, *11*, 190–202.

Cawsey, A. (1992). *Explanation and interaction*. Cambridge, MA: MIT Press.

Chang, R.C., Evens, M., Michael, J., & Rovick, A. (1994). Surface generation in tutorial dialogues based on a sublanguage study. *Proceedings of ICAST '94* (pp. 113–119).

Chi, M.T.H., De Leeuw, N.D., Chiu, M.H., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science*, *18*, 439–477.

Chi, M., Siler, S.A., Jeong, H., Yamauchi, T., & Hausmann, R.G. (2002). Learning from human tutoring. *Cognitive Science*, *25*, 471–533.

Cho, B.I., Michael, J.A., Rovick, A.A., & Evens, M. (1999). A curriculum planning model for an intelligent tutoring system. *Proceedings of the Florida Artificial Intelligence Symposium* (pp. 197–201).

Cohen, P.A., Kulik, J.A., & Kulik, C.C. (1982). Educational outcomes of tutoring: A meta-analysis of findings. *American Educational Research Journal*, *19*, 237–248.

Collins, A., Warnock, E.H., & Passafiume, J.J. (1975). Analysis and synthesis of tutorial dialogues. In G.H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 9, pp. 49–87). New York: Academic Press.

Conati, C., & VanLehn, K. (1999). Teaching meta-cognitive skills: Implementation and evaluation of a tutoring system to guide self-explanation while learning from examples. In S.P. Lajoie & M. Vivet (Eds.), *Proceedings of AIED-99* (pp. 297–304). Amsterdam: IOS Press.

Dawson-Saunders, B., Feltovich, P.J., Coulson, R.L., & Steward, D.E. (1990). A survey of medical school teachers identify basic biomedical concepts medical students should understand. *Academic Medicine*, *65*, 448–454.

Edelson, D. (1996). Learning from cases and questions: The Socratic case-based teaching architecture. *Journal of the Learning Sciences*, *5*, 357–420.

Elmi, M.A., & Evens, M.W. (1998). Spelling correction using context. *COLING-98* (pp. 360–364). Montreal.

Evens, M.W., Brandle, S.S., Chang, R.C., Freedman, R., Glass, M., Lee, Y.H., Shim, L.S., Woo, C.W., Zhang, Y., Zhou, Y., Michael, J.A., & Rovick, A.A. (2001). CIRCSIM-Tutor: An Intelligent Tutoring System Using Natural Language Dialogue. *Proceedings of the Midwest Artificial Intelligence and Cognitive Science Conference* (pp. 16–23). Oxford, OH.

Freedman, R. (2000a). Plan-based dialogue management in a physics tutor. *Proceedings of the Sixth Applied Natural Language Processing Conference* (pp. 52–59). Seattle, WA.

Freedman, R. (2000b). Using a reactive planner as the basis for a dialogue agent. *Proceedings of Florida Artificial Intelligence Research Symposium 2000* (pp. 203–208). Orlando, FL.

Freedman, R., Penstein Rosé, C., Ringenberg, M., & VanLehn, K. (2000). ITS tools for natural language dialogue: A domain-independent parser and planner. In G. Gauthier, C. Frasson, & K. VanLehn (Eds.), *Proceedings of the ITS 2000* (pp. 433–442), Montreal, LNCS 1839. Berlin: Springer.

Gertner, A., Conati, C., & VanLehn, K. (1998). Procedural help in Andes: Generating hints using a Bayesian network student model. *Proceedings of AAAI-98* (pp. 106–111). Madison, WI.

Glass, M. (1999). *Broadening input understanding in a language-based intelligent tutoring system*. Unpublished doctoral dissertation, Illinois Institute of Technology, Chicago.

Glass, M. (2000). Processing language input in the Circsim-Tutor intelligent tutoring system. *AAAI Fall Symposium on Building Dialogue Systems for Applications* (pp. 74–79).

Glass, M. (2001). Processing language input for an intelligent tutoring system. J.D. Moore, C.L. Redfield, & W.L. Johnson (Eds.), *Proceedings of Artificial Intelligence in Education* (pp. 210–221). Amsterdam: IOS Press.

Glass, M., Kim, J.H., Evens, M., Michael, J.A., & Rovick, A.A. (1999). Novice vs. expert tutors: A comparison of style. *Proceedings of the Midwest Artificial Intelligence and Cognitive Science Symposium* (pp. 43–49).

Graesser, A.C., Person, N.K., & Magliano, J.P. (1995). Collaborative dialogue patterns in naturalistic one-to-one tutoring. *Applied Cognitive Psychology*, *9*, 495–522.

Hume, G., Michael, J., Rovick, A., & Evens, M. (1996). Hinting as a tactic in one-on-one tutoring. *Journal of the Learning Sciences*, *5*, 23–47.

Khuwaja, R.A., Evens, M.W., Rovick, A.A., & Michael, J.A. (1992). Knowledge representation for an intelligent tutoring system based on a multilevel causal model. In C. Frasson, G. Gauthier, & G.I. McCalla (Eds.), *Intelligent tutoring systems*. *Proceedings of the Second International Conference*, *ITS '92* (pp. 217–224). Montreal, Canada. Berlin: Springer.

Kim, J.H., Glass, M., Freedman, R., & Evens, M. (2000). Learning the use of discourse markers in tutorial dialogue for an intelligent tutoring system. *Proceedings of Cognitive Science 2000* (pp. 262–267). Philadelphia, PA.

Kim, J.H., Freedman, R., Glass, M., & Evens, M.W. (in preparation). Annotation of tutorial goals for natural language generation.

Kuipers, B.J., & Kassirer, J.P. (1984). Causal reasoning in medicine: Analysis of a protocol. *Cognitive Science*, *8*, 363–385.

Li, J., Seu, J., Evens, M., Michael, J., & Rovick, A. (1992). Computer dialogue system (CDS): A system for capturing computer-mediated dialogues. *Behavior Research Methods, Instruments, and Computers (Journal of the Psychonomic Society)*, *24*, 535–540.

Michael, J.A. (1998). Students' misconceptions about perceived physiological responses. *American Journal of Physiology*, *274* (*Advances in Physiology Education*, *19*), S90–S98.

Michael, J.A., & Rovick, A.A. (1999). *Problem solving in physiology.* Upper Saddle River, NJ: Prentice-Hall.

Moore, J.D. (1995). *Participating in explanatory dialogues*. Cambridge, MA: MIT Press.

Moore, J.D., Lemaire, B., & Rosenblum, J. (1996). Discourse generation for instructional applications: Identifying and using prior relevant explanations. *Journal of the Learning Sciences*, *5*, 49–94.

Person, N., Bautista, L., Graesser, A.C., Mathews, E., & the Tutoring Research Group. (2001). *Proceedings of Artificial Intelligence and Education* (pp. 286–293). San Antonio, TX.

Rovick, A.A., & Michael, J.A. (1986). CIRCSIM: An IBM-PC computer teaching exercise on blood pressure regulation. *Proceedings of the International Union of Physiological Sciences 30th Conference* (p. 318). Vancouver, Canada.

Rovick, A.A., & Michael, J.A. (1992). The prediction table: A tool for assessing students' knowledge. *American Journal of Physiology*, *263 (Advances in Physiology Education, 8)*, S33–S36.

Sanders, G. (1995). *Generation of explanations and multi-turn discourse structures in tutorial dialogue based on transcript analysis*. Unpublished doctoral dissertation, Illinois Institute of Technology, Chicago.

Seu, J., Chang, R.-C., Li. J., Evens, M., Michael, J., & Rovick, A. (1991). Language differences in face-to-face and keyboard-to-keyboard tutoring sessions. *Proceedings of the 13th Annual Conference of the Cognitive Science Society* (pp. 576–580). Chicago, IL.

Shah, F., Evens, M., Michael, J.A., & Rovick, A.A. (2002). Classifying student initiatives and tutor responses in human keyboard-to-keyboard tutoring sessions. *Discourse Processes*, *33*, 23–52.

Siegel, S., & Castellan, N.J., Jr. (1988). *Nonparametric statistics for the behavioral sciences*. Boston: McGraw-Hill.

Stevens, A.L., & Collins, A. (1980). Multiple conceptual models of a complex system. In R. Snow, P. Frederico, & W. Montague (Eds.), *Aptitude, learning, and instruction* (Vol. II, pp. 177–197). Hillsdale, NJ: Erlbaum.

Stevens, A., Collins, A., & Goldin, S.E. (1982). Misconceptions in students' understanding. In D. Sleeman & J.S. Brown (Eds.), *Intelligent tutoring systems* (pp. 13–24). London: Academic Press.

Thompson, B.H. (1980). Linguistic analysis of natural language communication with computers. *Proceedings of COLING 80*, Tokyo, 190–201.

VanLehn, K., Freedman, R., Jordan, P., Murray, C., Osan, R., Ringenberg, M., Rosé, C.P., Schulze, K., Shelby, R., Treacy, D., Weinstein, A., & Wintersgill, M. (2000). Fading and deepening: The next steps for Andes and other model-tracing tutors. In G. Gauthier, C. Frasson, & K. VanLehn (Eds.), *Proceedings of the Intelligent Tutoring Systems 2000* (pp. 474–483). Montreal, LNCS 1839. Berlin: Springer.

Woo, C.W. (1992). *A multi-level dynamic instructional planner for an intelligent tutoring system*. ONR Technical Report. Computer Science Department, Illinois Institute of Technology.

Woolf, B. (1984). *Context-dependent planning in a machine tutor.* Unpublished Ph.D. dissertation, Department of Computer and Information Science, University of Massachusetts at Amherst. COINS Technical Report, 84–21.

Zhang, Y., Evens, M., Michael, J.A., & Rovick, A.A. (1990). Extending a knowledge base to support explanations. *Proceedings of the Third IEEE Conference on Computer-Based Medical Systems* (pp. 259–266).

Zhou, Y., Freedman, R., Glass, M., Michael, J.A., Rovick, A.A., & Evens, M. (1999a). What should the tutor do when the student cannot answer a question? *Proceedings of the Florida Artificial Intelligence Symposium* (pp. 187–191).

Zhou, Y., Freedman, R., Glass, M., Michael, J.A., Rovick, A.A., & Evens, M. (1999b). Delivering hints in a dialogue-based intelligent tutoring system. *Proceedings of AAAI '99* (pp. 128–134).

Zhou, Y., & Evens, M. (1999). A practical student model in an intelligent tutoring system. *Proceedings of the 11th IEEE International Conference on Tools with Artificial Intelligence* (pp. 13–18).

APPENDIX: STUDENT SURVEY

YOUR VIEWS ON CIRCSIM-TUTOR

1 = definitely yes ..2..3..4..5 = definitely NO

1.	**The print in the display was readable.**	**1 2 3 4 5**
2.	**The screen layout was helpful.**	**1 2 3 4 5**
3.	**The sequence of displays was appropriate.**	**1 2 3 4 5**
4.	**The system was easy to use.**	**1 2 3 4 5**
5.	**The introductory screens were helpful.**	**1 2 3 4 5**
6.	**The system's dialogue seemed varied and interesting.**	**1 2 3 4 5**
7.	**The tutor's hints and explanations were informative.**	**1 2 3 4 5**
8.	**I would prefer that the system always tell me about my mistakes immediately.**	**1 2 3 4 5**
9.	**CIRCSIM-Tutor helped me to understand the behavior of the baroreceptor reflex.**	**1 2 3 4 5**
10.	**CIRCSIM-Tutor improved my ability to predict the cardio-vascular responses to disturbances in blood pressure.**	**1 2 3 4 5**

Please comment on any of the preceding questions or any other issue (problems using the system, help that wasn't provided but should have been, changes you would like to see in the system).

(You may continue on the back if you have more to say.)