A Comparative Evaluation of Socratic versus Didactic Tutoring

Carolyn Penstein Rosé (rosecp@pitt.edu)
LRDC, Univ. of Pittsburgh, Pittsburgh PA, 15260, USA

Johanna D. Moore (jmoore@cogsci.ed.ac.uk)
HCRC, Univ. of Edinburgh, Edinburgh EH8 9LW, UK

Kurt VanLehn (vanlehn@pitt.edu)
LRDC, Univ. of Pittsburgh, Pittsburgh PA, 15260, USA

David Allbritton (dallbrit@condor.depaul.edu)
Dept. of Psychology, DePaul Univ., Chicago, IL 60614, USA

Abstract

While the effectiveness of one-on-one human tutoring has been well established, a great deal of controversy surrounds the issue of which features of tutorial dialogue separate effective uses of dialogue in tutoring from those that are less effective. In this paper we present a formal comparison of Socratic versus Didactic style tutoring that argues in favor of the Socratic tutoring style.

Introduction

Comparative studies of student learning have already demonstrated that one-on-one human tutoring is more effective than other modes of instruction. Tutoring raises students' proficiency as measured by pre and post-tests by a minimum of 0.40 standard deviations with peer tutors (Cohen et al., 1982), and to up to 2.0 standard deviations with experienced tutors (Bloom, 1984). One prominent component of effective human tutoring is collaborative dialogue between student and tutor (Fox, 1993; Graesser et al., 1995; Merrill et al., 1992). Nevertheless, a great deal of controversy surrounds the issue of which features of tutorial dialogue distinguish effective uses of dialogue in tutoring from those that are less effective.

In this paper we present the results of a formal evaluation of the relative effectiveness of Socratic versus Didactic tutoring in a simulated problem solving environment in the Basic Electricity and Electronics domain. In this study, the Socratic tutoring style is characterized by an emphasis on eliciting information from students through a directed line of reasoning. Thus, the tutor endeavors as much as possible to avoid giving information away. In contrast, in the Didactic tutoring style, the tutor begins extended interactions with students by presenting the student with an explanation of the material the student is meant to learn in the interaction. After the initial explanation, the tutor leads the student through a directed line of reasoning similar to that used in the Socratic condition, except that the questioning plays more of a role of drawing the student's attention to information that the tutor has already explained, rather than eliciting this information from the student. Since drawing the student's attention to the points already articulated by the tutor requires less from the students, the Didactic interactions tended to be significantly shorter than the Socratic interactions. They contained only 70% as many open ended questions and had more of a lecture like flavor than the Socratic interactions.

In a classroom instructional context, it has been well argued that receptive learning, i.e., by means of lectures, can be just as effective as discovery based learning provided that students have the requisite prior knowledge to learn the presented material meaningfully rather than by rote (Ausubel, 1978). However, this view has not been universally accepted by educational psychologists (Piaget, 1973; Vygotsky, 1978). Furthermore, one key difference between tutoring and classroom learning is that there is much more continuity and regularity in classroom learning. In contrast, tutoring is sporadic and decontextualized. Thus, within a classroom setting the teacher is much more familiar with the students' prior knowledge, and is in fact in a position to ensure that students are prepared to learn the material that is presented each day by arranging lessons to build one upon another. We argue that because this is not the case in a tutoring context, it is more critical to draw out the student's thought process in order to tailor the presentation of material to the students' needs.

Previous studies have argued the effectiveness of Socratic and other similar tutoring approaches. Recent research on student self-explanations supports the view that when students explain their thinking out loud it enhances their learning (Chi et al., 1989, Chi et al., 1994, Renkl, submitted). Students learn more effectively when they are given the opportunity to discover knowledge for themselves (Brown and Kane, 1988; Lovett, 1992; Pressley et al., 1992). Collins and Stevens (1982) report that the best teachers tend to use a Socratic tutoring style. A tutoring system based on the Collins and Stevens model (Wong et al., 1998) has received favorable reviews although it has not yet been subjected to a formal comparative evaluation.

Nevertheless, other studies have argued the effectiveness of Didactic style tutoring. In a recent study in which students read previously solved probability problems (Renkl, submitted), an experimental group that had the option to request further tutor explanation performed better than a control group that did not have that option, with an effect size of .5 sigma. Albacete (1999) found that students who received Didactic conceptual minilessons when they made errors learned more than students who received immediate flag feedback and could

request first a pointing hint, then a correct answer. Similarly, McKendree (1990) found that feedback messages with high content caused more learning than feedback messages with low content. Finally, it is reported in (Graesser et al., 1995) that ordinary human tutors seldom use Socratic tutoring, and yet are quite effective.

The results of our formal comparison presented here demonstrate a trend in favor of Socratic style tutoring over Didactic style tutoring.

Experimental Setup

The context of our work is a web-based course on basic electricity and electronics (BE&E). The system was developed with the VIVIDS authoring tool (Munro, 1994) at the Navy Personnel Research and Development Center in San Diego, CA. The original BE&E tutor was designed as a tool for classroom instruction. As we are interested in one-on-one tutoring, we observed how students interacted with the system and with a human tutor in a Wizard-of-Oz setup. The curriculum used in our experiments and prototype system consists of four lessons and six labs covering basic concepts of current, voltage, resistance, power, making measurements with a multimeter, and doing some simple computations and problem solving. Each lesson consists of three sections of between 10 and 25 pages of instructional text and graphical illustrations displayed in a Netscape window, as in Figure 1. After each section, the student was tested with between 3 and 5 multiple choice progress check questions, which act as a springboard for interaction with the tutor to address deficiencies in the student's understanding. After each lesson, the student was presented with one or two labs designed to test and reinforce the concepts introduced in the lesson. The students completed the labs by interacting with a simulated electronics workbench through a point-and-click interface, as in Figure 2.

The 37 subjects who participated in our data collection experiment were University of Pittsburgh undergraduates with little or no prior study of electricity or electronics. Each subject participated in two sessions, each of which lasted between two and two and a half hours. The tutor was a post-doc working at the Learning Research and Development Center at the University of Pittsburgh with some tutoring experience. While the student interacted with the system, the video signal to the student's monitor was split so that a tutor sitting behind a partition could watch the student's progress. The student and tutor had access to a chat interface that allowed them to type messages to each other. Although students sometimes initiated dialogues themselves, the majority of dialogues occurred when the tutor initiated a dialogue because the student either incorrectly answered an important progress check question or showed evidence of not being able to proceed with a lab.

Seventeen of the students participated in a pilot study, which we used to determine how much material on average that students are able to cover in the allotted time (five hours maximum) and which concepts from the domain come up most frequently in the tutoring dialogues.

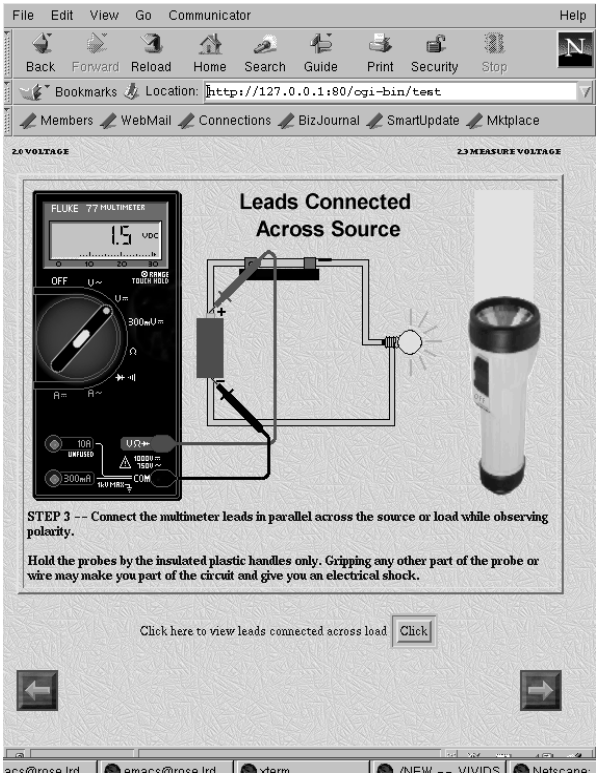

Figure 1: **Netscape Window**

Thus, we designed the pre/post-test for the formal study to focus on these troublesome concepts. We also slightly shortened the lessons by removing some material not essential to learning these concepts (e.g., how to interpret the colored stripes on resistors).

Twenty students took part in the final data collection effort. Each student was randomly assigned to either the Socratic condition or the Didactic condition, described above. Note that the Socratic tutoring style in this study did not typically include heavy use of Socratic irony (i.e., proof by contradiction) as a teaching tool.

Examples of Socratic and Didactic tutoring dialogues from our corpus are displayed in Figures 4 and 5 respectively. Both dialogues occurred during a lab in which the students were expected to find three places in a DC circuit where they could get a non-zero voltage reading. A very typical error students made was to attempt to get a non-zero voltage reading across a closed switch. Both the Socratic and Didactic dialogue examples occurred in response to this error. The important piece of information the tutor wanted to get across to the student in both cases was that it is only possible to get a non-zero voltage reading where there is a difference in charge, and there is no difference in charge across a switch. Notice that in both cases the tutor presented the students with questions to encourage the students to think through the reasoning behind their incorrect action. However, in the Socratic example, the student is doing more of the talking,

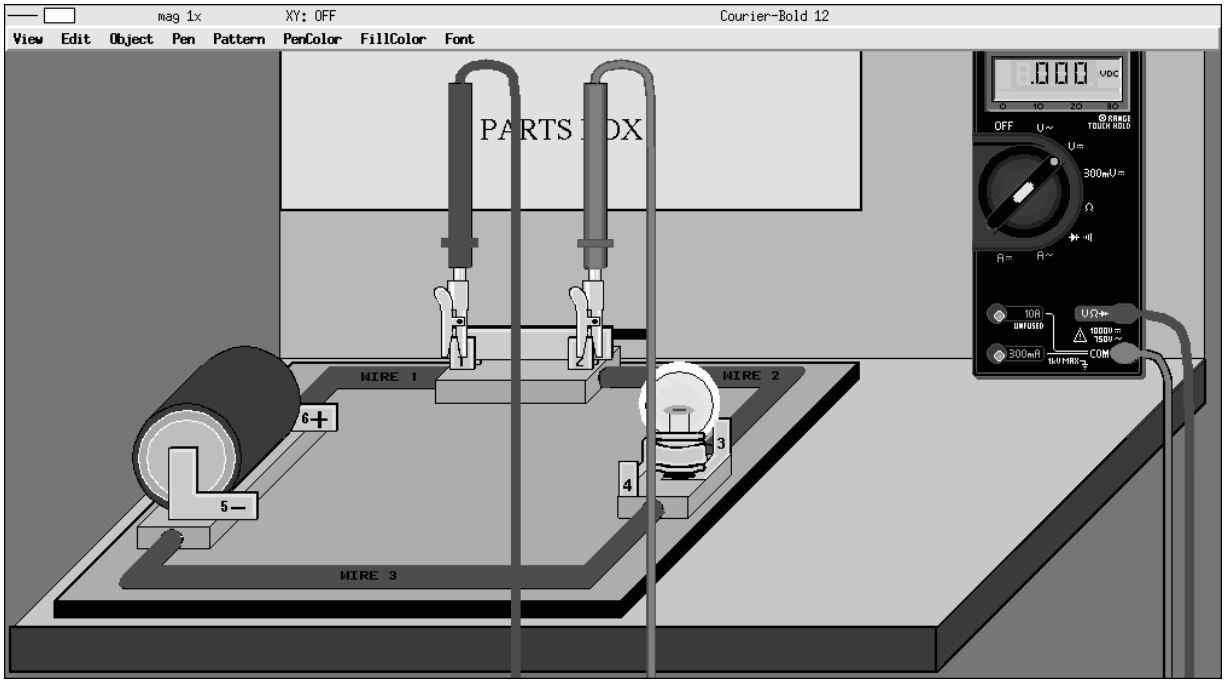

Figure 2: **Simulation Window**

where the Didactic example contains much more lecturing on the tutor's part. In the Didactic example, the tutor explained all that the student needed to know before asking the student any contentful questions. In contrast, in the Socratic example, the tutor asked contentful though questions from the beginning and explained as little as possible. In general, because in the Didactic condition the tutor was able to refer to parts of her own explanation that initiated the dialogue, the Didactic dialogues contained approximately 70% as many open ended questions such as why and how questions. Because in the Socratic condition the students were responsible for articulating as many key concepts as possible themselves, these open ended questions were essential for drawing out the students' reasoning.

Data Analysis

The data collected from the twenty students who participated in the formal study was used in a rule gain analysis to determine the relative effectiveness of the two alternative tutoring strategies. As in (VanLehn et al., 1998), the purpose of our rule gain analysis was to compute a correlation between student learning and the distinctive features of the two alternative tutorial styles. For each student we kept detailed records of their participation in our study as well as other information relevant to evaluating their learning. In particular, we recorded which of the two conditions the student was assigned to, the student's math and verbal SAT scores, the amount of time spent on each lesson and lab, the full text from all tutorial dialogues, and a pre and post-test. The same test was used for the pre-test as for the post-test. Note that students in

```
You can only get a non-zero voltage
reading across a portion of a circuit
where there is a difference in charge
(voltage potential)

There is a difference in voltage
potential across a source

There is a difference in voltage
potential across a load

Sources cause current to move because of
a difference in potential between their
two ends

Loads have a difference in potential
between their two ends because their
resistance causes them to convert
electrical energy into energy of a
different form
```

Figure 3: **Sample Rules for Difference in Voltage Potential**

the Didactic condition spent slightly more time interacting with the tutor and system overall (Socratic students spent an average of 192.68 minutes, SD=37.0; Didactic students spent an average of 198.72 minutes, SD=47.19), however the difference was not significant ($t(18)=.318$, $p=.754$).

For analyzing student learning, we codified the concepts we wanted to teach students into a list of 47 rules. For example, the five rules expressing basic concepts about voltage potential and voltage measurements are

```
Tutor:     So now let's think about the switch.  Is it a source or a load?

Student:   source

Tutor:     Why do you say that?

Student:   when it is turned off, no energy flows

Tutor:     OK, that is true.  But does it actually cause electricity to flow?
           Or is it merely letting what current was produced by the source
           flow through it to the rest of the circuit?

Student:   it does not create energy

Tutor:     Good.  So it is not a source.  Would you say it is a load then?

Student:   no

Tutor:     Good, it is neither a source nor a load.  So does it have a
           difference in charge between its two ends?

Student:   no

Tutor:     Very good.  So then can you measure voltage across a switch?

Student:   no

Tutor:     Right.  So how can you modify the way you have the leads hooked up
           now so that there will be a difference in charge between where the
           two leads are hooked up?

Student:   since i know there is a third place to hook them up, is it possible
           to connect the one to the negative side of the battery and connect
           the other to the positive of the light bulb?

Tutor:     Yes, that would be perfect.  Go ahead and do that.
```

Figure 4: Sample Socratic Dialogue

found in Figure 3. The pre/post test consists of 40 questions that require the students to use the rules either by applying a rule in order to solve a problem, directly articulating a rule, drawing a diagram to illustrate a rule, answering a question about a rule, or recognizing a diagram illustrating a rule. We coded each question in our pre/post test with a list of rules that applied to solving that problem and how those rules applied. For each rule we then developed a formula to assess the student's level of mastery of that rule based on which of the questions where the rule applied the student answered correctly. The formulas designed for computing the mastery score for each rule were based on the assumption that different ways rules apply to questions give different amounts of evidence about how well students have mastered the corresponding rule. For example, directly articulating a rule gives more evidence of student knowledge than answering a multiple choice question. Solving a problem by using a rule gives an even stronger indication. A total mastery score for each student was computed for both the pre and post-test by summing the mastery scores for the 47 individual rules. We also computed a gain score for each student by counting the number of rules for which each student demonstrated a higher mastery score on the post-test as compared with that on the pre-test. For each rule we noted how many students in each condition achieved a higher mastery score for that rule on the post-test as compared to the pre-test.

An ANCOVA with condition as the independent variable, pre-test score as the covariate, and post-test score as the dependent variable confirmed a trend for students in the Socratic condition to learn more (pre-test mean = 10.41, pre-test SD = 7.5, post-test mean = 27.54, post-test SD = 7.28) than students in the Didactic condition (pre-test mean = 14.29, pre-test SD = 53.12, post-test mean = 25.5, post-test SD = 6.31), $F(1, 18) = 3.13$; $p < .1$. Interestingly, despite the fact that the students in the Socratic condition had a lower average pre-test score than the students in the Didactic condition, they achieved a higher average post-test score. The effect size (mean Socratic gain score - mean Didactic gain score / SD Didactic gain score) was 1 sigma. Additionally, for each rule we computed a chi-square to determine whether the number of students who demonstrated learning on that rule in the Socratic condition was significantly higher than the number of students who demonstrated learning on that rule in the Didactic condition. The difference was only significant ($p < .05$) for two rules although for every rule more students in the Socratic condition than in the Didactic condition demonstrated learning. The probability that more students in the Socratic condition would demon-

```
Tutor:      Do you remember in the lesson that it said that every point on a
            conductor is electrically the same?

Student:    yes.

Tutor:      Good.  That means that there is no difference in potential energy
            between one point on the conductor and another point.  So, if both
            leads are attached to the same conductor, there is no difference
            in potential (in other words, no difference in charge) and thus no
            force to measure.  What you need is to have there be a difference
            in potential between where the red lead is attached and where the
            black lead is attached.  This is achieved whenever voltage is
            "created" as in a battery or "used up" as in a light bulb.  So,
            where do you suppose you could attach the leads now to achieve
            that?

Student:    to the light bulb, it seems to be my only source

Tutor:      Right.
```

Figure 5: Sample Didactic Dialogue

strate learning for all 47 rules by chance is very small, specifically $.5^{47}$. We found during our analysis, however, that in spite of having randomly assigned students to conditions, the average SAT score for students in the Socratic condition (1161, SD = 208) was marginally higher than that for students in the Didactic condition (961, SD = 192), F (1, 15) = 4.27, $p < .06$. Finding even a marginal trend with only 20 subjects suggests that the effect may be real even if the statistical power is not sufficient.

A re-analysis in which gain score (post test score minus pretest score) was the dependent variable and SAT score was a covariate was conducted for the 17 participants for whom SAT scores were available. The effect of condition was not significant in this analysis, F (1, 14) = 1.64, p ¿ .20, although the loss of power resulting from the exclusion of three participants from this analysis may have contributed to the lack of statistical significance. Thus, although the trend for greater gains in the Socratic condition was still evident after controlling for SAT scores, we can not conclusively rule out the influence of this possible confound on the results.

Next we checked for an aptitude-treatment interaction by dividing the students into two subsets, with students having above the mean SAT scores in one subset and those with below the mean SAT scores in the other subset. When the ANCOVA was performed on each subset separately, a trend was demonstrated for students in the Socratic condition to perform better within each subset. For above average SAT students, F (1,7) = 1.77, $p < .23$. For below average SAT students, F (1,4) = 5.62, $p < .1$. This seems to be a result of the fact that the below average SAT students tended to have uniformly low pre-test scores and variable post-test scores whereas the above the average SAT students had variable pre-test scores and uniformly high post-test scores.

Discussion and Current Directions

In this paper we present a study in which we explore the relative effectiveness of two alternative tutoring styles, which we have referred to as Socratic versus Didactic. The purpose of our study was to explore alternative methods of encouraging knowledge construction via tutorial dialogue in order to determine how to use dialogue most effectively to this end. The results of our rule gain analysis demonstrate that students in the Socratic condition learned more effectively than students in the Didactic condition, although more data collection is necessary in order to verify the level of statistical significance.

Based on our findings, we are building a dialogue-enhanced version of the original BE&E tutoring system. A small prototype dialogue based version has already been built covering the portion of our curriculum concerned with teaching about measuring voltage in DC circuits (Rosé et al., 1999). Note that in the example Socratic dialogue in Figure 4, the tutor affirmed what was correct in the student's response, that when a circuit is turned off no energy flows, and then addressed specifically what was lacking in the student's explanation. The same ability to evaluate the content of student explanations is required for the tutor to determine when it is no longer beneficial to continue prompting a student with leading questions. This type of sensitivity in tutor response is only possible in an intelligent tutoring system when the system can understand what the student says. Thus, a major focus of our work has been on robust natural language understanding (Rosé, 2000). Our robust core understanding component for English is currently being integrated into three different tutoring systems and is available for use on other projects[1].

[1] Parties interested in obtaining the Atlas core understanding component should contact Carolyn Rosé at `rosecp@pitt.edu`.

References

Albacete, P. L. (1999). *An Intelligent Tutoring System for Teaching Fundamental Physics Concepts*. PhD thesis, University of Pittsburgh, Pittsburgh, PA.

Ausubel, D. (1978), *Educational Psychology: A Cognitive View*, Holt, Rinehart and Winston, Inc.

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

Brown, A. L. and Kane, M. J. (1988). Preschool children can learn to transfer: Learning to learn and learning from example. *Cognitive Psychology*, 20:493–523.

Chi, M. T. H., Bassok, M., Lewis, M. W., Reimann, P., and Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science*, 13(2):145–182.

Chi, M. T. H., de Leeuw, N., Chiu, M. H., and LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science*, 18(3):439–477.

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

Collins, A. and Stevens, A. (1982). Goals and methods for inquiry teachers. In Glaser, R., editor, *Advances in Instructional Psychology, Vol. 2*. NJ: Lawrence Erlbaum Associates, Hillsdale.

Fox, B. A. (1993). *The Human Tutorial Dialogue Project: Issues in the design of instructional systems.* Lawrence Erlbaum Associates, Hillsdale, New Jersey.

Graesser, A. C., Person, N. K., and Magliano, J. P. (1995). Collaborative dialogue patterns in Naturalistic One-to-One Tutoring. *Applied Cognitive Psychology*, 9:495–522.

Lovett, M. C. (1992). Learning by problem solving versus by examples: The benefits of generating and receiving information. In *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society*. NJ:Erlbaum.

McKendree, J. (1990). Effective feedback content for tutoring complex skills. *Human-Computer Interaction*, 5:381–413.

Merrill, D. C., Reiser, B. J., and Landes, S. (1992). Human tutoring: Pedagogical strategies and learning outcomes. Paper presented at the annual meeting of the American Educational Research Association.

Munro, A. (1994). Authoring interactive graphical models. In de Jong, T., Towne, D. M., and Spada, H., editors, *The Use of Computer Models for Explication, Analysis and Experiential Learning*. Springer Verlag.

Piaget, J. (1973). To understand is to invent. New York: Grossman.Resnick, L. B. (1989). Developing mathematical knowledge. AmericanPsychologist, 44, 162-169.

Pressley, M., Wood, E., Woloshyn, V. E., Martin, V., King, A., and Menke, D. (1992). Encouraging mindful use of prior knowledge: Attempting to construct explanatory answers facilitates learning. *Educational Psychologist*, 27:91–109.

Reiser, B. J., Copen, W. A., Ranney, M., Hamid, A., and Kimberg, D. Y. (in press). Cognitive and motivational consequences of tutoring and discovery learning. In *Cognition and Instruction*.

Renkl, A. (submitted). Worked-out examples: Instructional explanations support learning by self-explanations.

Rosé, C. P. (2000). A framework for robust semantic interpretation. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*.

Rosé, C. P., Di Eugenio, B., and Moore, J. D. (1999). A dialogue based tutoring system for basic electricity and electronics. In *Proceedings of the Ninth World Conference on Artificial Intelligence in Education*.

VanLehn, K., Siler, S., and Baggett, W. (1998). What makes a tutorial event effective. In *Proceedings of the Twenty-first Annual Conference of the Cognitive Science Society*.

VanLehn, K., Siler, S., Murray, C., Yamauchi, T., and Baggett, W. B. (in press). Human tutoring: Why do only some events cause learning? cognition and instruction.

Vygotsky, L. S. (1978). Mind in society: The development of higherpsychological processes (M. Cole, V. John-Steiner, S. Scribner, & ESouberman, Eds.). Cambridge, MA: Harvard University Press.

Wong, L. H., Quek, C., and Looi, C. K. (1998). TAP-2: A framework for an inquiry dialogue-based tutoring system. *International Journal of Artificial Intelligence in Education*, 9.

Acknowledgments

This research is supported by the Office of Naval Research, Cognitive and Neural Sciences Division (Grants N00014-91-J-1694 and N00014-93-I-0812) and NSF Grant 9720359 to CIRCLE, a center for research on intelligent tutoring.