Evaluating the Effectiveness of a Cognitive Tutor for Fundamental Physics Concepts

Patricia L. Albacete (albacete@isp.pitt.edu)
Intelligent Systems Program; 607 Dixie Drive
Pittsburgh, PA 15235 USA

Kurt A. VanLehn (Vanlehn@cs.pitt.edu)
Learning, Research and Development Center; University of Pittsburgh
Pittsburgh, PA 15260 USA

Abstract

In this article we describe and analyze the evaluation of the Conceptual Helper, an intelligent tutoring system that uses a unique cognitive approach to teaching qualitative physics. The results of the evaluation are encouraging and suggest that the proposed methodology can be effective in performing its task.

Introduction

Several studies (e.g. Hake, 1998; Halloun & Hestenes, 1985a, 1985b) have revealed that solving physics problems of a qualitative nature, such as the one presented in figure 1, pose a great cognitive challenge for most students taking elementary mechanics classes. They uncover naïve conceptions that are seldom removed or modified while completing their courses. Several attempts have been made to improve this situation though none has met with great success (Hake, 1998). Given that mechanics is a required course for most science majors, there is a clear need to improve its instruction. Toward this end we developed an intelligent tutoring system called the Conceptual Helper that follows a cognitive teaching strategy which is deployed emulating effective human tutoring techniques as well as successful pedagogical techniques and less cognitive demanding methods (Albacete, 1999; Albacete & VanLehn, 2000). In this article we describe the evaluation of the system and discuss its implications.

Two steel balls, one of which weights twice as much as the other, roll off of a horizontal table with the same speeds. In this situation:

a) both balls impact the floor at approximately the same horizontal distance from the base of the table.
b) the heavier ball impacts the floor closer to the base of the table than does the lighter.
c) the lighter ball impacts the floor closer to the base of the table than does the heavier.

Figure 1. Example of a qualitative problem

Brief description of the Conceptual Helper

The Conceptual Helper is an intelligent tutoring system (ITS) designed to coach students through physics homework problem solving of a qualitative nature, i.e., those problems that do not require the use algebraic manipulation to be solved but so require the application of conceptual knowledge. The tutor is basically a model-tracing ITS enhanced by the use of probabilistic assessment to guide the remediation. As a model-tracing ITS it contains a cognitive model that is capable of correctly solving any problem assigned to the student. Model tracing consists of matching every problem-solving action taken by the student with the steps of the expert's solution model of the problem being solved. This matching is used as the basis for providing immediate feedback to students as they progress through the problem. The system also has a student model which is represented by a Bayesian network. Each node in the network represents a piece of conceptual knowledge that the student is expected to learn or a misconception that the tutor can help remedy. Each node has a number attached to it that indicates the probability that the student will apply the piece of knowledge when it is applicable. As the student solves a problem, the probabilities are updated according to the actions taken by the student.

The challenge for the tutor is to decide when to intervene and what to say when it does so. This task is particularly challenging in this domain because tutoring of qualitative knowledge usually takes the form of verbal discussions, which given the state of the art of natural language processing is not an option for the computer tutor. To take care of the issue of when to intervene, we emulated human tutors in two ways: first, by giving immediate feedback (red for incorrect; green for correct) on each student entry (Merrill et al., 1992) and second, by helping the student with post-problem reflection (Katz & Lesgold, 1994; Katz et al., 1996). However, most of our work went into the second issue—deciding what to say when intervening. Novel approaches were developed in three areas: 1) the teaching strategy, 2) the manner in which the knowledge is deployed, and 3) the way in which misconceptions are handled.

The Conceptual Helper's teaching strategy

Several studies (e.g. Van Heuvelen, 1991) have characterized students' knowledge of conceptual physics as a collection of ill-structured, unconnected facts and concepts which remain almost the same after completion of their physics classes. In contrast, cognitive science theory describes experts' knowledge bases as being well structured and *highly connected* (e.g. Chi & Koeske, 1983). Based on these findings, the teaching strategy embedded in the Conceptual Helper tries to make students' knowledge bases akin to the experts' by concentrating on teaching students the *links* that connect the domain's concepts of interest rather than the concepts in themselves.

The word "links" has been traditionally used in Semantic Networks to describe two-place predicates such as "is-a" or "part-of". However, we use the word "links" to describe rich qualitative rules that integrate pieces of knowledge. The links that the Conceptual Helper focuses can be inferred from the principles or from the definitions of the concepts of the domain. For example, one of the target links is "the direction of the net force applied to an object is the same as the direction of the object's acceleration." This connection between the concept of acceleration and the concept of net force can be inferred from Newton's second law. Likewise, the link "if the acceleration of an object is zero, then the object's velocity is constant" can be inferred from the definition of the concept of acceleration. These types of links are not evident to the students, in the sense that, even if students can repeat without hesitation the definition of acceleration and Newton's second law, by and large, they are generally not able to assert the links between concepts that follow from those definitions (Reif, 1995). However, these types of links are essential for reasoning qualitatively about the motion of objects and for solving the qualitative problems.

How is the target knowledge taught?

The knowledge presented by the teaching strategy is deployed using a combination of: a) effective tutoring techniques, such as hinting through dialogues (Fox, 1993; Lepper et al., 1990), b) successful pedagogical techniques, like the use of molecular view of matter (Murray et al., 1990), and c) less cognitive demanding methods, such as using anthropomorphism (diSessa, 1993; Roschelle, 1992) and objects belonging to the material ontology (Chi, 1992) to reify abstract physics concepts. Figure 2 describes a mini-lesson that the tutor would present to the student when explaining the link "if (in a linear motion) the velocity of an object is decreasing, then the object's velocity and its acceleration have opposite directions." It exemplifies some of the techniques used by the tutor.

The manner in which misconceptions are handled

To help students replace their misconceptions with scientifically correct knowledge, the Conceptual Helper presents students with the basic line of reasoning underlying the correct interpretation of the phenomena that are the base of the misconception. This is as opposed to using discovery environments or computer-simulated experiments, which are two common ways in which teachers have tried to correct misconceptions (Hake, 1998). We believe that it is not setting up the (simulated) equipment, making the runs, recording the data, and inducing a pattern that convinces a student of a certain piece of knowledge, but rather the line of argument itself. Knowing the correct line of reasoning enables the student to self-explain the phenomenon, which has been argued (Chi, 1996) to be an effective means for learning.

Evaluation of the Conceptual Helper

Forty-two students taking Introductory Mechanics classes were recruited and randomly divided into a Control group and an Experimental group. Both groups took a paper-and-pencil pre-test that consisted of 29 qualitative problems, 15 of which belonged to the Force Concept Inventory test[1]. Then they solved some problems with the Andes system receiving appropriate feedback according to the group they belonged to. The students in the Control Group had their input turned green or red depending on the correctness of the entry. Then, in the case of an incorrect action, the students could ask for help by making a choice from a help menu. The kind of help they received consisted of simple hints such as "the direction of the vector is incorrect." If the student asked for more help, they would just be told the correct answer. On the other hand the students in the experimental group received the green/red feedback depending on whether their action was correct but when the input was incorrect the Conceptual Helper intervened as explained above. After the students finished solving the problems with the system they took a post-test which was the same as the pre-test with the exception of a few changes in the cover stories of some problems. Among the problems included in the pre-test, post-test, and Andes there were multiple-choice questions and problems that required an explicit solution. Finally the students were asked to complete a questionnaire expressing their evaluation of the system.

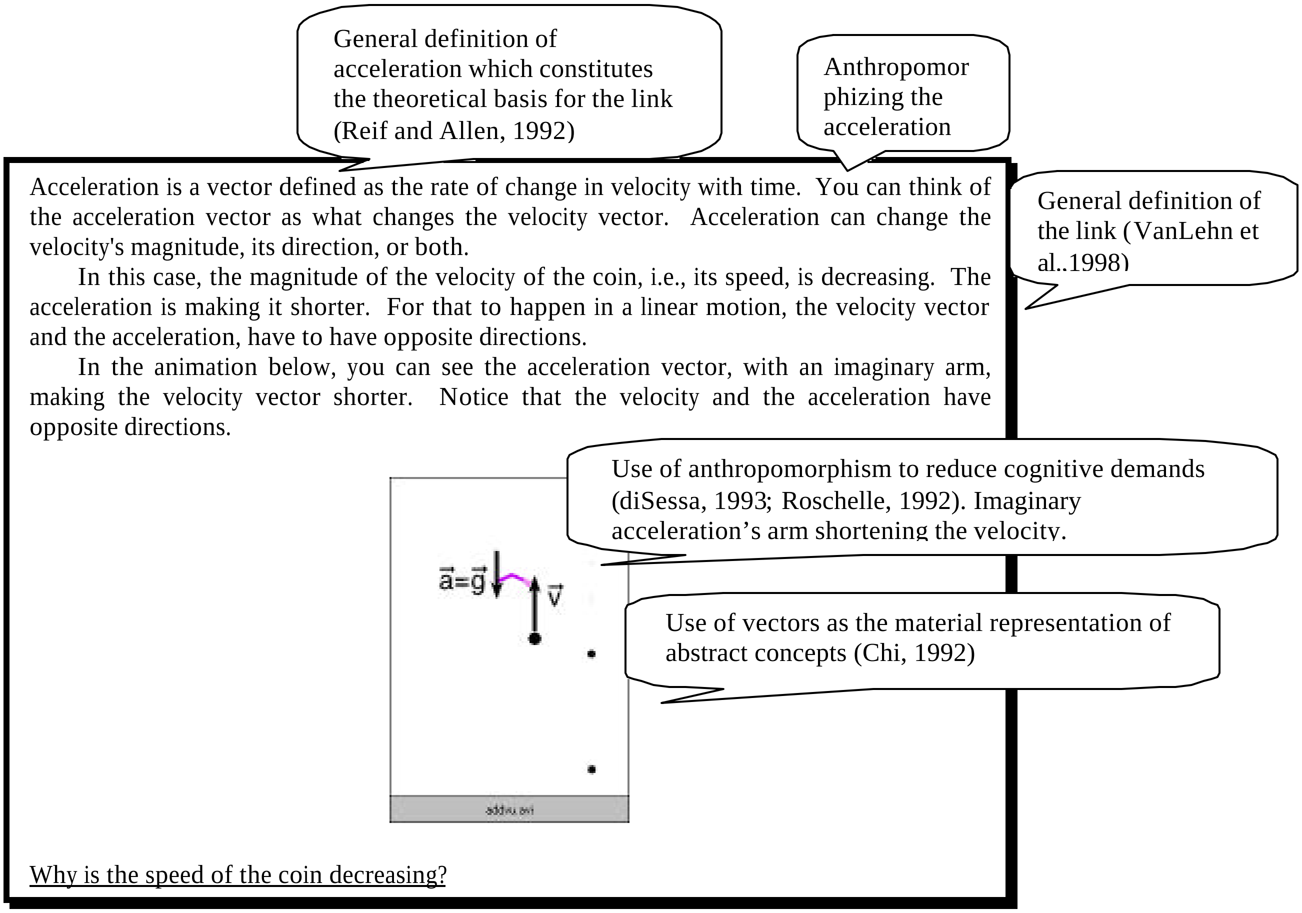
Results and their interpretation

The data gathered in such a way was analyzed in different ways.

1. <u>T-test using the gain scores from pre-test to post-test as the dependent measure</u>

Before comparing the gains of the two groups, we first checked whether their initial competencies were equivalent. The mean pretest score of the control group was 33.7 with standard deviation of 7.47. The mean pretest score of the experimental group was 31.36 with a standard deviation of 8.14. No reliable difference was found between the two groups (t(40)=0.965, p=0.34). Next, the gain scores from pre-test to post-test were compared. The mean of the control group was 4.12 with a standard deviation of 5.33.

[1] The Force Concept Inventory Test has become the standard test across the US to measure conceptual understanding of elementary mechanics (Hakes, 1998).



General definition of acceleration which constitutes the theoretical basis for the link (Reif and Allen, 1992)

Anthropomorphizing the acceleration

General definition of the link (VanLehn et al..1998)

Acceleration is a vector defined as the rate of change in velocity with time. You can think of the acceleration vector as what changes the velocity vector. Acceleration can change the velocity's magnitude, its direction, or both.

In this case, the magnitude of the velocity of the coin, i.e., its speed, is decreasing. The acceleration is making it shorter. For that to happen in a linear motion, the velocity vector and the acceleration, have to have opposite directions.

In the animation below, you can see the acceleration vector, with an imaginary arm, making the velocity vector shorter. Notice that the velocity and the acceleration have opposite directions.

Use of anthropomorphism to reduce cognitive demands (diSessa, 1993; Roschelle, 1992). Imaginary acceleration's arm shortening the velocity.

Use of vectors as the material representation of abstract concepts (Chi, 1992)

Why is the speed of the coin decreasing?

Figure 2: example of a mini-lessons

The mean of the experimental group was 7.47 with a standard deviation of 5.03. A reliable difference was found (t(40)=2.094, p=0.043, two-tailed). This statistically significant difference suggests that the intervention of the Conceptual Helper had a positive impact on the students' understanding of the concepts as well as on their ability to abandon common misconceptions.

2. Effect size

Effect size is a standard way to compare the results of one pedagogical experiment to another. One way to calculate effect size, used in Bloom (1984) and many other studies, is to subtract the mean of the gain scores of the control group from the mean of the gain scores of the experimental group, and divide by the standard deviation of the gain scores of the control condition. That calculation yields (7.47-4.12/5.33 = 0.63). This result was comparable with peer and cross-age remedial tutoring (effect size of 0.4 according to Cohen, Kulik and Kulik, 1982). Some better results have been obtained with interventions that lasted a whole semester or academic year. For example, Bloom (1984) found an effect size of 2.0 for adult tutoring in replacement of classroom instruction and Anderson et al. (1995) reported an effect size of 1.0 for their tutoring systems. However, our results were achieved with only two hours of instruction.

3. The fraction of the maximum possible gain realized (G)

Another measure that is used in the literature to compare the results of the FCI test is G = (Sf - Si) / (100 - Si), where Si and Sf are the pre- and post-test scores in percent (Hake, 1998). The nationwide score on the FCI test for traditionally taught classes is G = 0.25. For classes that are taught in a more interactive manner, G is between 0.36 and 0.68 (Mazur, 1997). The results obtained considering all the problems were the following: The mean of the control group was 0.26 with a standard deviation of 0.36. The mean of the experimental group was 0.43 with a standard deviation of 0.25[2]. The mean G for the control group matches that for traditionally taught classes. However, the G for the experimental group, 0.43, places it with the classes that are taught in a more interactive manner.

[2] Even though in the literature G is reported for each particular classroom in which a teaching method is applied and no statistical comparisons are made, we performed a two-tailed t-test to compare the G of the control and experimental group. The results were t(40)=1.84, p=0.073.

4. Existence of an aptitude-treatment interaction (ATI)

Innovative interventions sometimes cause higher gains for students with higher pre-test scores. What we want to find, of course, is that students with lower pre-test scores improved more, as they are the students who need more help. In order to see whether there was an aptitude treatment interaction (ATI) and which way it would go, the experimental group was divided into two groups according to whether the student's pre-test score was above or below the median. The mean gain of the low pre-test score group was 10.68 with a standard deviation of 5.00. The mean gain of the high pre-test score group was 4.27 with a standard deviation of 2.39. A statistically significant difference between the gain scores was found (t(20)=3.83, p=0.002, two-tailed).

A similar analysis was done with the control group. The results obtained were as follows: The mean gain of the low pre-test score group was 5.90 with a standard deviation of 5.87. The mean gain of the high pre-test score group was 2.35 with a standard deviation of 4.31. No statistically significant difference between the mean of the gain scores was found (t(18)=1.54, p=0.14, two-tailed). Figure 3 illustrates the results for the experimental and control groups.

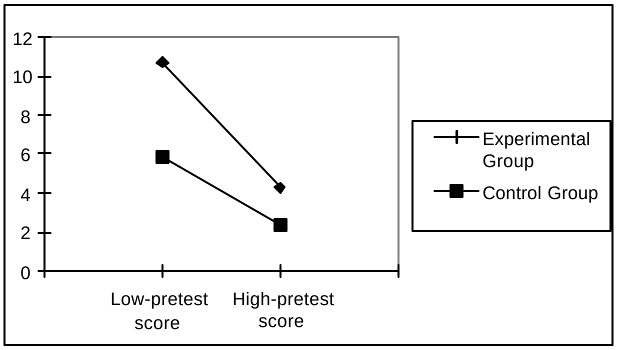


Figure 3. Mean gain of the low- and high- pretest score groups in the experimental and control groups.

It was encouraging to find that in the experimental group the poorer subjects' knowledge gains were significantly higher than those of good students, revealing that there was a desirable ATI. Additionally, it should be noted that the lower gain score in the high pre-test score group was not a consequence of a ceiling effect. The mean pre-test score of the high-pretest group was 38.04 with a standard deviation of 4.32. Since the maximum score is 49 there was an opportunity for this group to have a gain score very close to that achieved by the group of poorer students. Moreover, one student got a post-test score of 49, which indicates that the post-test did not require unlearnable knowledge.

5. Detailed analysis of the individual pieces of knowledge and the effectiveness of each mini-lesson

A more detailed analysis was performed with the objective of determining the effectiveness of each mini-lesson in conveying the appropriate pieces of knowledge and in fostering their transfer. The method used basically consisted of comparing whether receiving a mini-lesson had an effect on gaining versus not gaining the knowledge. Gaining the knowledge means giving an incorrect answer in the pre-test and a correct one in the post-test. Not gaining the knowledge means giving an incorrect answer in both pre- and post- tests. In the case were the target knowledge was addressed during explicit problem solving (e.g. for the rule "if an object's velocity is constant then its acceleration is zero") only students from the experimental group were considered, because only they could receive the mini-lessons. In the case were the target knowledge was addressed only through multiple-choice questions (e.g. the rule "heavier/lighter objects fall faster"), we compared the gains of the experimental group to the gains of the control group. The reason for doing this is that all the students in the experimental group received these mini-lessons, which were presented whenever the student answered a multiple-choice question (whether correctly or incorrectly). In the cases where the knowledge was addressed in both explicit and multiple-choice questions (e.g. the rule "force that continues to act after no contact"), we investigated whether receiving a mini-lesson during explicit problem solving would have any effect on gaining the rule. Hence only students from the experimental group were considered.

Statistical power problems prevented the analysis of most of the 18 target rules from showing a reliable relationship between receiving a mini-lesson and gaining the piece of knowledge. For some rules, almost all students received the corresponding mini-lesson, whereas for other rules, too few students received the mini-lesson. Nonetheless, there were a few rules where the relationship between mini-lessons and gain could be tested. They are described in Table 1. In all cases a Fisher's exact test (Hayes, 1994) was performed.

In most cases shown in Table 1 the number of students in each group was not large enough to provide statistical power, even if all those who received the mini-lesson gained and all those who did not receive the mini-lesson failed to gain (see third row of Table 1). Nonetheless, the data suggest a positive relationship between receiving a mini-lesson and gaining the corresponding knowledge.

6. Summary of students' comments about the system

Students were asked to fill out a short questionnaire to express their opinion about the system. The rating of the different aspects of the system was done on a scale ranging from 1 to 5 where 5 was the best possible score. Students gave as score of 4 or above to all different aspects of the system (e.g., explanations that are clear to understand) which show a favorable acceptance of the system as well as a fairly high degree of liking of the mini-lessons.

Discussion

The evaluation of the tutor suggests that the teaching strategy followed by the Conceptual Helper along with its methodology for deploying the target knowledge and handling misconceptions, is effective in accomplishing the task it was designed to perform. The experimental group surpassed the control group in every statistical test performed. Moreover, a detailed examination of the

Table 1. Relationship between receiving a mini-lesson and gaining knowledge for selected rules

Rule name	Group	Gainers	Non-gainers	Total	P	P for most extreme cases
Influence of weight on horizontal motion	**Experimental**	12 (.44)	2 (.07)	14 (.52)	0.005	<0.005
	Control	4 (.15)	9 (.33)	13 (.48)		
	Total	16 (.59)	11 (.41)	27		
When the velocity is constant the acceleration is zero	**Got mini-lesson**	5 (.5)	0 (0)	5 (.5)	0.08	0.08
	Did not get mini-lesson	2 (.2)	3 (.3)	5 (.5)		
	Total	7 (.7)	3 (.3)	10 (1)		
Heavier/lighter objects fall faster	**Experimental**	3 (.6)	0	3 (.6)	0.1	0.1
	Control	0	2 (.4)	2 (.4)		
	Total	3 (.6)	2 (.4)	5		
Vertical motion takes over horizontal motion	**Experimental**	9 (.45)	2 (.1)	11 (.55)	0.38	P<0.05 if all students in Exp. group are gainers
	Control	6 (.3)	3 (.15)	9 (.45)		
	Total	15 (.75)	5 (.25)	20		
Force that continues to act after no contact	**Got mini-lesson**	5 (.42)	3 (.25)	8 (.67)	0.24	P<0.05 all students that did not get mini-lesson were non-gainers
	Did not get mini-lesson	1 (.08)	3 (.25)	4 (.33)		
	Total	6 (.5)	6 (.5)	12 (1)		

The numbers in parenthesis represent the proportions with respect to the grand population

effectiveness of each individual mini-lesson showed a trend in favor of using the lesson.

Several studies (e.g. Halloun & Hestenes, 1985a) suggest that practice on solving quantitative problems does not transfer to conceptual problem solving. For instance, student who get full marks in their physics course still score poorly on the FCI test. On the other hand, elaborate confrontation-based, interactive instruction (e.g. see Hake, 1998) does raise scores on the FCI test, and by approximately the same amount as the Conceptual Helper. We believe that both forms of instruction are successful, at least in part, for the same reasons.

First, both handle misconceptions and errors with a form of confrontation. Both present students with situations (problems) and ask them to express their reasoning while solving them. In the Conceptual Helper, they do that by either taking an action, such as drawing a force, in an explicit solution problem or by choosing an answer in a multiple-choice question. If the action taken is incorrect, they are confronted with their erroneous knowledge by getting a mini-lesson. In interactive instruction the confrontations are quite elaborate and often involve doing experiments (e.g., Hake, 1992, McDermott, Shaffer & Somers, 1994, or White, 1993). What is interesting is that the evaluation of our system suggests that, in the case of misconceptions, confrontation based on *simply showing the correct line of reasoning* to describe the phenomena under consideration can be just as effective in remediating misconceptions as the more elaborate, time-consuming kinds traditionally used to teach conceptual physics. Additionally, the evaluation suggests that, for correcting conceptual errors (or lack of knowledge), confrontation based on teaching the links that connect the concepts of the domain in the manner presented by the Conceptual Helper, may help the students build a more organized and better connected knowledge base, which in turn may facilitate qualitative reasoning.

A second factor underlying the success of both forms of instruction is that they both use conceptual problems instead of quantitative problems. This facilitates transfer, but it does not make it trivial. In particular, the Conceptual Helper does not “teach to the test” i.e., it does not teach exactly what the students are tested on. For example, the last rule in Table 1, which corresponds to the common misconception that there exists a force in the direction of the motion that continues to act after an object has been set in motion, shows a trend in favor of receiving a mini-lesson. The mini-lesson was received by students when they made a mistake in solving a problem that dealt with describing the motion of a box sliding on a frictionless surface after it has been pushed. On the other hand, the post-test problem analyzed in Table 1 involved describing the forces acting on a ball thrown up in the air. Hence the situations presented in both problems were quite different even if the underling misconception involved was the same.

In summary, it seems that the Conceptual Helper is just as effective but more efficient than other forms of qualitative physics instruction, in part, possibly because both forms of instruction use conceptual problems and confrontation. The next step in this line of research is to develop efficient and effective methods for *integrating* conceptual and quantitative learning.

References

Albacete, P.L. (1999). An Intelligent Tutoring System for teaching fundamental physics concepts. *Unpublished doctoral dissertation*. Intelligent Systems Program, University of Pittsburgh. Pittsburgh, Pennsylvania.

Albacete, P.L. & VanLehn, K.A. (2000). (in press). *Fifth International Conference on Intelligent Tutoring Systems*. ITS'2000, Montreal, Canada.

Anderson, J.R., Corbett, A.T., Koedinger, K.R., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *The Journal of the Learning sciences*, 4(2) 167-207.

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.

Chi, M.T.H. (1992). Conceptual change within and across ontological categories. In Gier, R. (Ed.) *Cognitive models of science: Minnesota studies in the philosophy of science*. University of Minnesota Press, Minneapolis, MN.

Chi, M.T.H. (1996). Constructing Self-Explanations and Scaffolded Explanations in Tutoring. *Applied Cognitive Psychology*, 10, S33-S49.

Chi, M.T.H. & Koeske, R.D. (1983). Network Representation of a Child's Dinosaur Knowledge. *Developmental Psychology* 19(1), 29-39.

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.

diSessa, A.A. (1993). Toward an Epistemology of Physics. *Cognition and Instruction*, 1993, 10(2&3), 105-225.

Fox, B.A. (1993). *The Human Tutorial Dialogue Project: Issues in the Design of Instructional Systems*. Lawrence EribaumAssociates, Hillsdale, NJ.

Hake, R. R. (1992). Socratic Pedagogy in the Introductory Physics Lab, *Phys. Teach*. 30, 546 (1992).

Hake, R.R. (1998). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics*, 66(64).

Hayes, W. (1994). *Statistics*. Holt, Reinhart & Winston, Inc. 5th edition.

Halloun, I.A., & Hestenes, D. (1985a). The initial knowledge state of college physics students. *American journal of Physics* 53 (11) 1043-1055.

Halloun, I.A., & Hestenes, D. (1985b). Common sense knowledge about motion. *American journal of Physics* 53 (11) 1056-1065.

Katz S. & Lesgold A. (1994). Implementing Post-problem Reflection within Coached Practice Environments. In *Proceedings of the East-West International Conference on Computer Technologies in Education* (Part I; pp. 125-130). P. Brusilovsdy, S. Dikareva, J. Greer, V. Petrushin (Eds.). Crimea, Ukraine.

Katz, S., Lesgold, A., Eggan, G., Greenberg, L. (1996). Towards the Design of a More Effective Advisors for Learning by Doing Systems. *Proceedings of the Third International Conference on Intelligent Tutoring Systems*, ITS'96. Montreal, Canada. Springer-Verlag.

Mazur, E. (1997). *Peer Instruction*. Prentice Hall series in educational innovation. Upper Saddle River, NJ.

McDermott, L. C., Shaffer, P. S., & Somers M. D. (1994). Research as guide for teaching introductory mechanics: An illustration in the context of the Atwood's machine. *American Journal of Physics* 62 (1) 46-55.

Merrill, D.C., Reiser, B J., Ranney, M., & Trafton, J.G. (1992). Effective tutoring techniques: A comparison of human tutors and intelligent tutoring systems. *Journal of the Learning Sciences*, 2,2 77-306.

Murray, T., Schultz, K., Brown, D., & Clement, J. (1990). An Analogy-Based Computer Tutor for Remediating Physics Misconception. *Interactive Learning Environments* 1(2), 79-101.

Reif, F. (1995). Understanding and Teaching Important Scientific Thought Processes. *American Journal of Physics*, January 1995.

Reif, F. & Allen S. (1992). Cognition for Interpreting Scientific Concepts: A study of Acceleration. *Cognition and Instruction*, 9(1), 1-44.

Roschelle, J. (1992). Learning by Collaborating: Convergent Conceptual Change. *The Journal of the Learning Sciences*, 2(3), 235-276.

Lepper, M.R., Aspinwall, L., Mumme, D., & Chabay, R.W (1990). Self-perception and social perception processes in tutoring: subtle social control strategies of expert tutors. In J. M. Olson & M. P. Zanna (Eds.), *Self-inference processes: The Sixth Ontario Symposium in Social Psychology*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Van Heuvelen, A. (1991). Learning to think like a physicist: A review of research-based instructional strategies. *American Journal of Physics*, 59(10), 891-896.

VanLehn, K., Siler, S., Murray, C, & Bagget, W.B. (1998). What Makes a Tutorial Event Effective? *In Proceedings of the Twenty-first Annual Conference of the Cognitive Science Society*, Hillsdale, NJ: Erlbaum.

White, B.Y. (1993) ThinkerTools: Causal Models, Conceptual Change, and Science education. *Cognition and instruction*, 10(1), 1-100.

Acknowledgments

This research was supported by a grant from the Cognitive Science Division of ONR, N00014-96-1-0260.