# Computer-assisted assessment of student understanding in physics

Suzanne M. Lea
*Department of Mathematics, University of North Carolina at Greensboro, Greensboro, North Carolina 27412*

Beth A. Thacker and Eunsook Kim
*Department of Physics, The Ohio State University, Columbus, Ohio 43210*

Kathleen M. Miller
*Department of Physics, Miami University, Hamilton, Ohio 45011*

*(Received 10 February 1993; accepted 30 August 1993)*

A computer program that presents and grades physics problems is described. The reliability of the grading is discussed in the context of diagnostic tests. The concept of computerized testing in physics is shown to be feasible. It was found that graphics answer representations can be used to measure student understanding: students with correct graphics answers provide correct explanations and conversely. The sample problem studied did not produce any false negative results; the only false positive results were caused by incomplete explanations which are easily caught by keyword checking.

## INTRODUCTION

Tests are used for two purposes: to evaluate a student's knowledge in order to compare it with the student's previous knowledge or with the knowledge of other students; and to evaluate a student's knowledge in order to diagnose learning problems or misconceptions. Many teacher-developed and most standardized tests are short-answer tests: multiple choice, true-false, matching. These types predominate because they require the least time and labor to grade when large numbers of students are involved. However, a short-answer test is primarily suited for measuring factual recall. When short-answer tests attempt to measure problem solving skills or process, the severely constrained answer space both encourages memorization (of answers or algorithms for obtaining them) and limits the ability of the test to serve as a diagnostic tool. Furthermore, students may not acquire nor value skills such as analysis, synthesis, and evaluation of solutions, since these skills are not tested.

In physics, the ability to apply knowledge is more important than the ability to recall knowledge. Consequently, physics tests emphasize solving problems. Expert problem solving relies on a deep understanding of physics concepts and a hierarchical organization of knowledge.[1] Using tests to probe student conceptual understanding of physics and student knowledge organization and providing a diagnostic capability in tests requires either numerous staff to grade the tests and perform the diagnosis or technology. We have developed a computer program which accepts data for a test problem in the form of text files. Any problem can be used, as long as its answer can be represented in terms of requiring the user to move objects on the screen to particular spots. The program is referred to as a *testing engine*, in analogy with *inference engine* in expert systems work, to denote independence of the program from the particular problems used.

Previous work on using the computer in physics courses has focused on the computer as a laboratory and computational tool[2] or as a simulator and tutor.[3] McDermott and Trowbridge[4] discuss the computer and physics education research; Bork and Lochhead[5] and Mestre and Gerace[6] discuss its use in problem-solving research.

Although using the computer as a tutor requires assessing understanding, in many cases the assessment is done using standard (pencil and paper) means. Bork[2] used a variation of multiple-choice questions in computer testing. Chabay and Sherwood[7] use manipulation of objects on the computer screen to assist assessment of understanding in electricity.

## I. DEVELOPMENT OF THE ENGINE

There were three design goals for the project: test problems should be easy to add, the grade given by the computer should be correct, and the program should run on any common classroom computer. The first goal dictated that problem data be input as text files, requiring no programming effort on the part of the user. The second goal led to a decision to avoid input in English and to represent answers by the manipulation of objects on the computer screen (a graphics answer representation). The third goal led to the choice of the cT programming language,[8] which runs on both Macintosh and PC-compatible computers. This language also facilitates input of problem data as text files.

As a preliminary investigation of the feasibility of presenting problems via computer, we chose two problems where typical answers from students were available in the literature, so we could compare the answers students gave during use of the computer program with those given in interview situations. The two problems involved the location of the image formed by a plane mirror[9] and analysis of the motion of a modified Atwood's machine: a block on a table connected by a rope over a pulley to a hanging block.[10] The latter problem was animated to show the motion either in approximately real time or in slow motion.

The problems were administered via computer to volunteers, in-service teachers taking a physics course at UNCG, with no interviewer present. All interactions between users and the computer were logged to a data file. From the preliminary study, we drew the following conclusions.

(1) Any computerized testing program needs to ensure the representation of answers used is understood by the user. Some users had difficulty understanding the top-view diagrams used for the mirror problems.

(2) Providing feedback in the form of additional information about the problem when an answer is incorrect leads testwise users to change their answers regardless of their level of understanding. Hence, such feedback should not be provided.

(3) Tests should not be lengthy. Problems should be presented individually for best results.

(4) In any animated program, there is a strong "video game" effect: the user manipulates objects and watches the animation, but uses guesses rather than analysis to solve the problem.

(5) It is possible to present valid and reliable problems, problems which measure the same student abilities as the written versions and where the computer grade agrees with that given by a human.

(6) Designing problems for diagnostic testing requires knowing the misconceptions students have about the problem. Such problems must be based on research results to be successful.

## II. EVALUATING THE ENGINE

The aim is that the testing engine should not only grade problems reliably, but also discover where and why the student erred. This aim can be achieved only if problems can be devised which meet two criteria: (1) the explanation or solution of the written problem provides information allowing diagnosis of student errors; and (2) answers for the computer version of the problem can be constructed which give the same information as do written explanations or solutions. The engine was evaluated using problems from the *Physics by Inquiry*[11] materials. These materials contain very carefully developed diagnostic test problems designed to probe student understanding and elicit misconceptions held by the student. Hence, they meet the criterion that the answers should provide information allowing diagnosis of student errors.

The problem most extensively investigated, from the Electric Circuits module,[12] asks the student to use a model of electric current to predict bulb brightness in a dc electric circuit. Figure 1 shows the written version of the problem. Students in the course have not yet studied voltage; the brightness of a bulb is used to indicate qualitatively the amount of current passing through it. Although this problem does not appear difficult, many students in calculus-

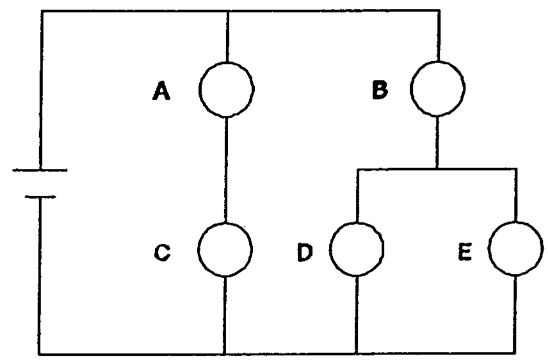
Experiment 4.8

A. Without setting up the circuit below, use your model for electric current to predict how the brightness of bulb A will compare to that of bulb B.

How will the brightness of bulb B compare to that of bulb D?

How will the brightness of bulb A compare to that of bulb C?

Explain your reasoning.




B. Set up the circuit and observe the actual brightness of the bulbs. Explain any differences between your predictions and your observations.

C. Check your reasoning with a staff member.

*Figure 1. The problem used to evaluate the testing engine. This problem is taken from the Electric Circuits module of the* Physics by Inquiry *materials.The test nature of the problem is indicated by part C. Batteries are assumed to be ideal and bulbs identical throughout the module.*

based physics courses cannot provide correct answers to similar problems.[13]

Figure 2 shows the first screen of the computer version of this diagnostic test. The student is asked to predict bulb brightness by moving appropriate icons to the blanks. The problem is basically multiple-choice; however, 6 possible icon choices for each of 5 blanks means there are $6^5$ or 7776 possible answers (of which 10 are correct, since we are interested only in relative order). In addition, the student must construct this answer instead of selecting it from a list.

After predicting brightness, the computer asks the student to connect the circuit and change predicted brightness to match observed brightness if necessary. (The problem is designed to be used in a laboratory setting.) The student then types in an explanation of bulb brightness. This not

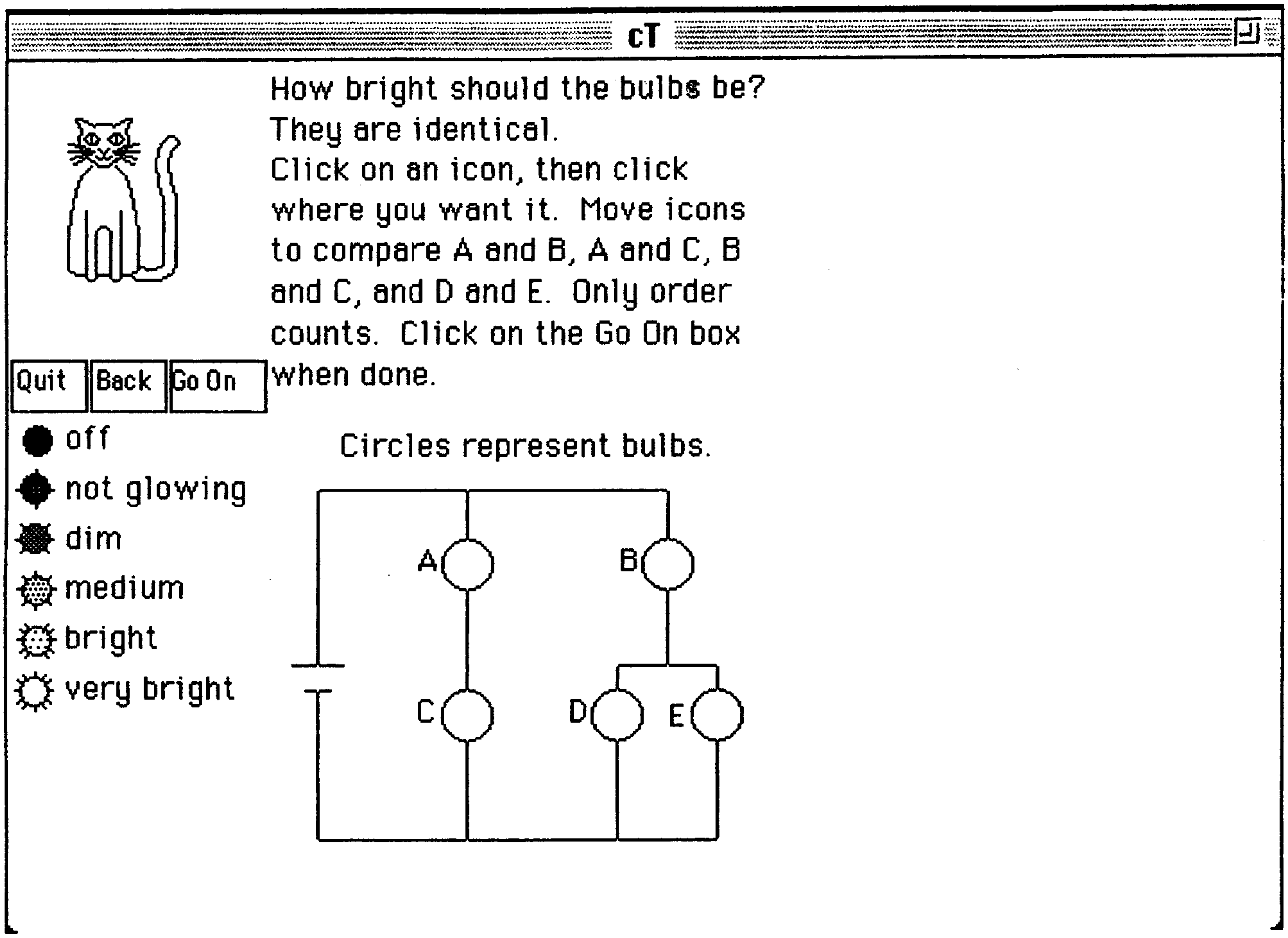



*Figure 2. The first screen of the computer version of the test problem. The student is asked to rank the brightness order of bulbs A and B, A and C, B and D, and D and E. Since students have not yet studied voltage, they are unable to provide a ranking for A and D from their model of current.*

only provides the student with the benefits of expressing her understanding in her own words,[14] but also provides us with a way of checking the reliability of the computer's grading of the test. (The typed explanation need not be required once the validity and reliability of the problem are assured.) All answers are recorded in a student data file for future reference.

Finally, the student is asked to indicate the relative magnitude of current at various points in the circuit (Fig. 3). For this part, there are $6^6$ or 46 656 possible answers, of which 4 are correct. Students have not yet used ammeters; therefore the current meter has been designed to look somewhat like a thermometer, a more familiar measuring instrument to most people. Students have had no difficulty accepting the instrument drawn as an instrument that measures current.

The validity of this problem is measured by whether it elicits the same misconceptions as the original and whether patterns of wrong answers provide useful diagnoses of student difficulties. Students' typed explanations indicate the problem is valid.

Reliability is measured by whether the computer gives the same grade as a human. The possible grades for a diagnostic test are right or wrong. The computer grade is based on the manipulation of screen objects; the human grade is based on the typed explanation. Reliability was tested with 54 students (in-service and preservice teachers) enrolled in physics classes using *Physics by Inquiry* at The University of North Carolina at Greensboro, at The Ohio State University, and at The University of Washington. The test population was determined by the fact that *Physics by Inquiry* materials are used primarily in classes for teachers. Since even calculus-based physics students find problems similar to this one difficult,[13] we would expect the same results with them as with our test population.

Results were classified into two types: the computer and the human judge agreed on the grade (49, or 91% of cases), and the computer and the human judge disagreed (5, or 9% of cases). Cases where the computer and human agreed included 16 students who failed to complete the test problem, 21 students who got the problem right, and 12 students who got the problem wrong. There were no cases where the computer thought the answer was wrong and the human thought the answer was correct (false negative judgments). There were 5 cases where the computer thought the answer was correct and the human judged the answer was wrong (false positive judgments). This type of judgment is particularly dangerous, since the student may proceed without a complete understanding. All 5 cases involved incomplete explanations: either the student did not explain the

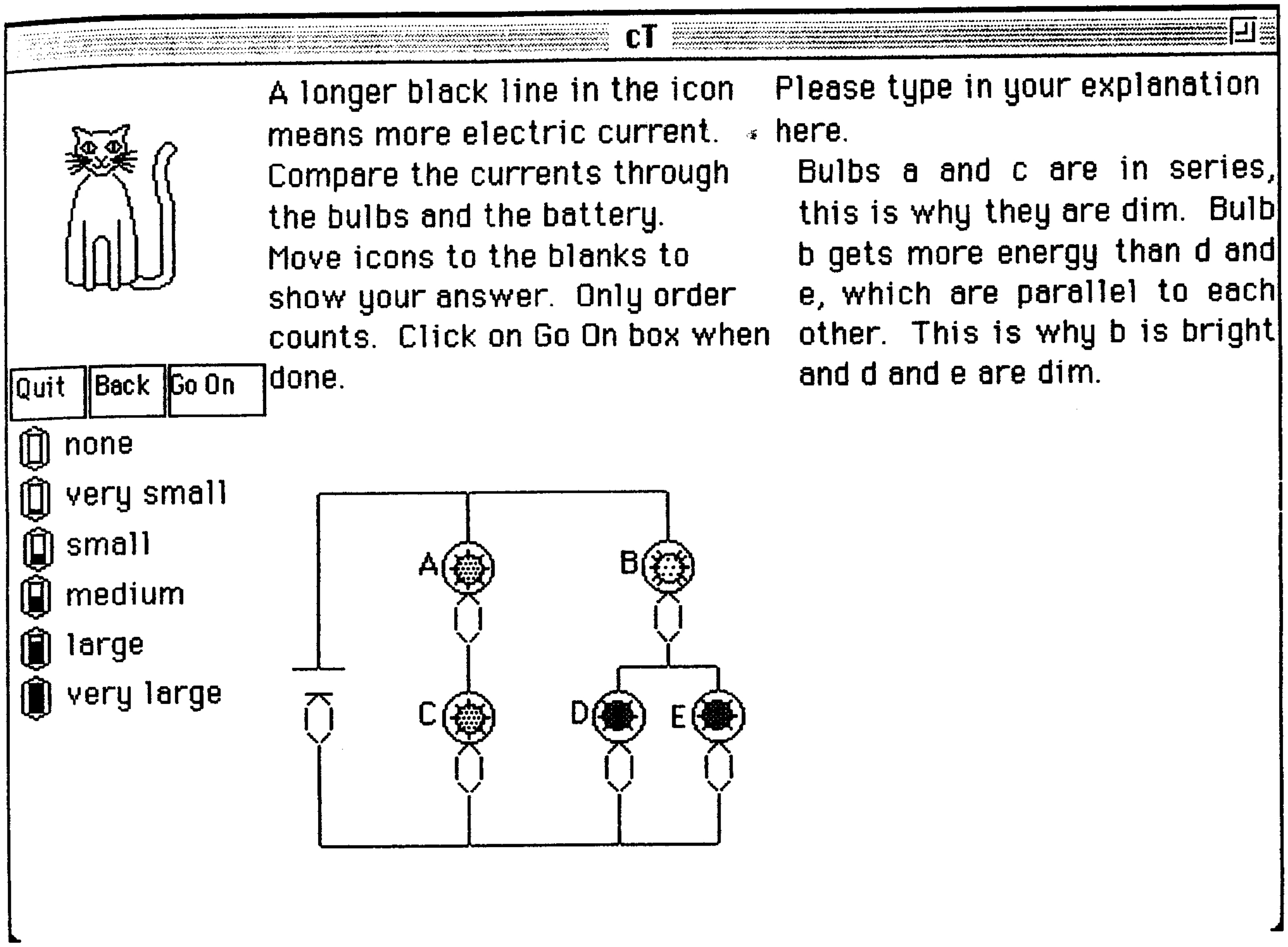



*Figure 3. The last screen of the computer version of the test problem. The student is asked to rank current values at various points in the circuit. Since they have observed the brightness of the bulbs in an actual circuit prior to answering this question, they are expected to provide a correct ranking of the currents through bulbs A and D, as well as the rest of the bulbs and the battery.*

brightness of all bulbs in the circuit (1 case) or the student explained brightness in terms of series and parallel connections instead of in terms of current and resistance (4 cases). The latter group of students may not understand the nature of scientific explanations.[1]

In these cases, the program gives more information than the written answer. To test whether the program grading is reliable requires the student to explain using a model of electric current. Consequently, we have incorporated a facility for keyword checking of the verbal explanation. The keyword checking uses data files containing lists of key words or phrases. To handle spelling or typographical errors in a primitive way, wild-card matching is allowed at the end of a word or phrase: for example, one can state that any word beginning *curr* provides a match for the word *current* we desire to see in the explanation. In the data file, the correct combinations of words or phrases that make the explanation complete are also specified. For example, one can specify that *flow* is a synonym for *current* by allowing the combination *current or flow*.

Figure 4 shows the keyword checking on an actual student explanation. The student has not used the terms *current* or *resistance*. For either type of incomplete explanation, the program provides a message and allows the student to edit or alter her explanation. With the keyword check, the program is 100% reliable for the sample tested. Of course, the keyword check does not catch complete but wrong explanations.

## III. CONCLUSIONS

We have shown the concept of computerized diagnostic testing in physics is feasible. We find that graphics answer representations can be used to measure student understanding: students with correct graphics answers provide correct explanations and conversely. The sample problem produced no false negative results; the only false positive results were caused by incomplete explanations which can easily be caught by keyword checking.

We are presently adding other problems to provide a more extensive test of the engine. We note that the validity of new problems must be assured before measuring reliability. In the future we plan to expand the number of types of answers supported by the engine and to test its reliability with other groups of students.

## ACKNOWLEDGMENTS

We thank L.C. McDermott and the members of the Physics Education Group at the University of Washington for help-

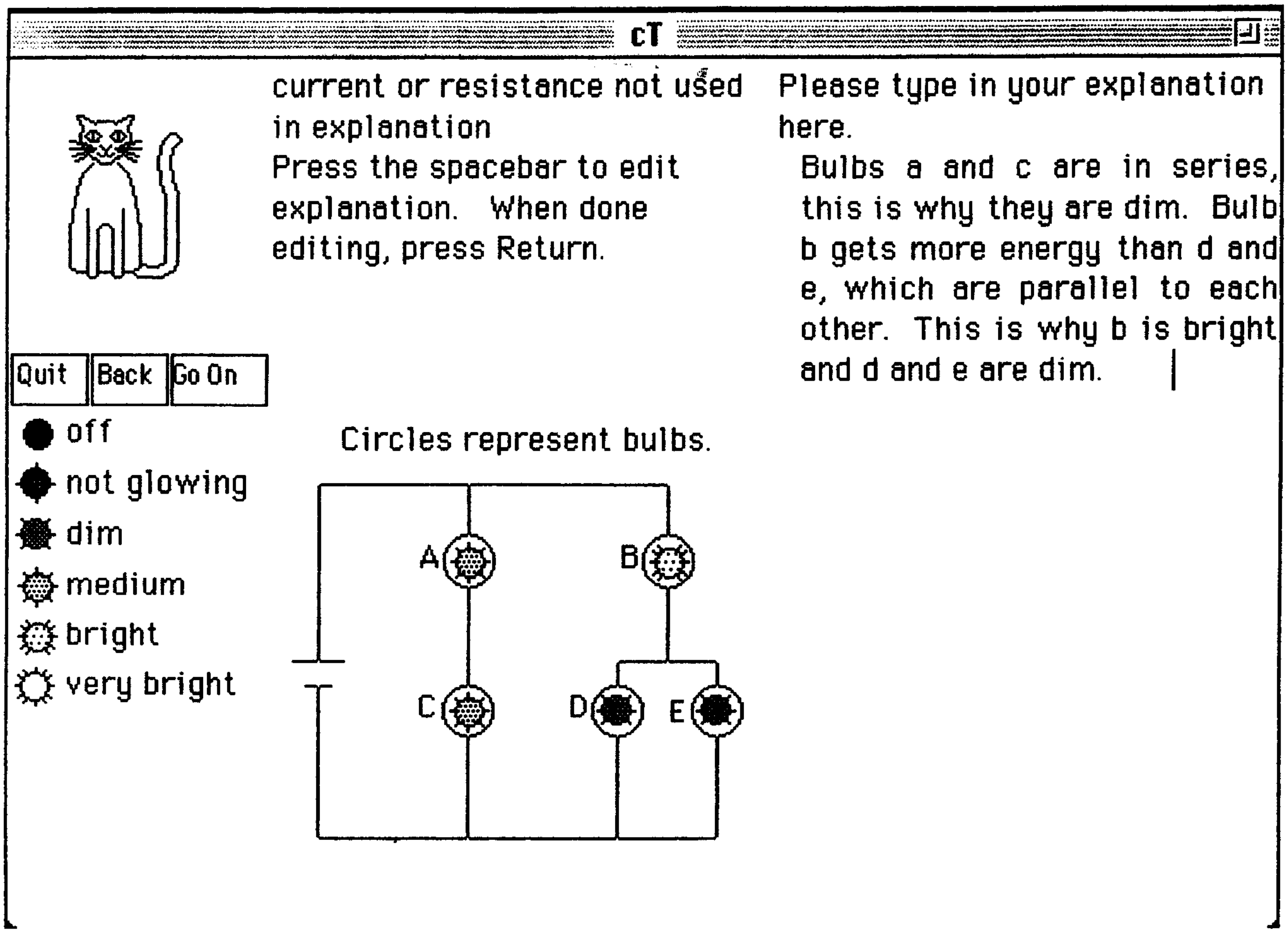



*Figure 4. Keyword checking for incomplete explanations. The student is informed of what is lacking and given the opportunity to alter or add to her explanation.*

ful conversations about diagnostic tests in the context of *Physics by Inquiry*. S.M.L. thanks D. Trowbridge for useful conversations about educational software, B. Sherwood for assistance with the cT language, and F. Reif for the suggestion that students whose explanations do not use a model may have a fundamental misunderstanding of the nature of scientific explanations. We thank E. L. Jossem for insisting on and assisting with friendly-user interfaces. S.M.L. acknowledges gratefully the hospitality of K. G. Wilson and The Ohio State University Physics Department, where much of this work was done.

## REFERENCES

1. F. Reif, J. Larkin, and G. C. Brackett, Am. J. Phys. **44**, 212 (1976); J. Larkin and F. Reif, Eur. J. Sci. Educ. **1**, 191 (1979); F. Reif, in *The Teaching of Problem-Solving in Engineering and Related Fields*, edited by J. L. Lubkin (American Society for Engineering Education, Washington, DC 1980); F. Reif, Phys. Teacher **19**, 310 (1981); F. Reif and J. I. Heller, Educ. Psychol. **17**, 102 (1982); B. S. Eylon and F. Reif, Cognition Instruct. **1**, 5 (1984); J. I. Heller and F. Reif, *ibid.* **1**, 177 (1984).
2. See, for example, S. E. Koonin, in *Computers in Physics Instruction*, edited by E. F. Redish and J. S. Risley (Addison-Wesley, Reading, MA 1990), pp. 329–337; P. W. Laws, *ibid.* pp. 22–32; R. K. Thornton, *ibid*, p. 177; P. W. Laws, Phys. Today **44**(12), 24 (1991); W. M. MacDonald, E. F. Redish, and J. M. Wilson, Comput. Phys. **2**, 23 (1988); C. Misner and P. Cooney, *Spreadsheet Physics* (Addison-Wesley, Reading, MA 1991); R. K. Thornton and D. R. Sokoloff, Am. J. Phys. **58**, 858 (1990).
3. See, for example, A. B. Arons, Science **224**, 1051 (1984); A. Bork, Phys. Today **34**(9), 24 (1981); B. Cabrera, in *Computers in Physics Instruction*, edited by E. F. Redish and J. S. Risley (Addison-Wesley, Reading, MA, 1990), pp. 77–83; R. W. Chabay, *ibid.* pp. 290–301; D. Trowbridge, *ibid.*, pp. 115–116; D. Trowbridge, R. W. Chabay, and B. A. Sherwood, *ibid.*, pp. 143–145. H. Gould and J. Tobochnik, *An Introduction to Computer Simulation Methods—Applications to Physical Systems*, Pt. I, II (Addison-Wesley, Reading, MA, 1988); E. Mazur, Comput. Phys. **5**, 31 (1991); G. E. Oberem, Proceedings of the Conference on Computers in Post-Secondary Education: Learning in Future Education, University of Calgary, Alberta, Canada, 1987, (unpublished) pp. 179–184; E. F. Taylor, Am. J. Phys. **57**, 508 (1989); D. Trowbridge and A. Bork, in *Computers in Education*, edited by R. Lewis and D. Tagg (North-Holland, New York, 1981), p. 325.
4. L. C. McDermott, Am. J. Phys. **58**, 452 (1990); D. E. Trowbridge, in *Computers in Physics Instruction*, edited by F. F. Redish and J. S. Risley (Addison-Wesley, Reading, MA, 1990), pp. 282–289.
5. A. Bork and J. Lochhead, Probl. Solv. **1**, 7 (1980).
6. J. Mestre and W. Gerace, *Proceedings of the 8th Annual Conference Cognitive Science Society* (Lawrence Erlbaum, Hillside, NJ, 1986), pp. 741–746.
7. R. W. Chabay and B. A. Sherwood, CDEC Technical Rep. No. 89-10, Center for Design of Educational Computing, Carnegie-Mellon University, Pittsburgh, PA 15213-3890.
8. B. A. Sherwood and J. N. Sherwood, *The cT Programming Language* (Falcon Software, Wentworth, NH, 1988).
9. F. M. Goldberg and L. C. McDermott, Phys. Teach. **24**, 472 (1986).
10. D. J. Grayson, "Use of the Computer for Research on Instruction and Student Understanding in Physics," Ph.D. thesis, University of Washington, 1990 (unpublished).

11. L. C. McDermott, Physics by Inquiry, Physics Education Group, Physics Department, University of Washington, Seattle, WA, 1982–1992 (unpublished).
12. L.C. McDermott and P. S. Shaffer, Am. J. Phys. **60**, 994 (1992); P. S. Shaffer and L. C. McDermott, Am. J. Phys. **60**, 1003 (1992).
13. A. B. Arons, Am. J. Phys. **50**, 13 (1982); L. C. McDermott and E. H. van Zee, in *Aspects of Understanding Electricity*, edited by R. Duit, W. Jung, and C. v. Rhoeneck (Verlag Schmidt & Klaunig, Kiel, Germany, 1984), pp. 39–48; B. Thacker, E. Kim, K. Trefz, and S. M. Lea (in preparation).
14. A. B. Arons, *A Guide to Introductory Physics Teaching* (Wiley, New York, 1990).