Behavior Research Methods, Instruments, & Computers
2004, 36 (2), 234-240

The DIAGNOSER project: Combining assessment and learning

ANNE THISSEN-ROE and EARL HUNT
University of Washington, Seattle, Washington

and

JIM MINSTRELL
Facet Innovations, Seattle, Washington

DIAGNOSER is an Internet-based tool for classroom instruction. It delivers continuous formative assessment and feedback to high school physics students and their teachers about the correct and incorrect concepts and ideas the students may hold regarding physical situations. That is, it diagnoses misconceptions that underlie wrong answers of students, such as a confusion of velocity with acceleration. We use data about patterns of student responses, particularly consistency of errors from question to question, to improve the system's understanding of student concepts.

Learning benefits from rapid, well-directed feedback. This is well known. Unfortunately, the further that a student advances in school, the harder it is to obtain such feedback. A high school science teacher will deal with upward of 100 students a day, so it is virtually impossible for the teacher to spend very much time with any one student. The situation is even worse at the university level, where feedback may occur only on midterms and end-of-term examinations. Technological solutions to this problem have been proposed, ranging from Skinner's *teaching machines* to modern-day courses on the World-Wide Web. With the exception of Skinner's work, which was rooted in his theories of behaviorism, most of these systems are driven by a desire to deliver content and correct error, rather than being built around a psychological theory. Also, most of the teaching programs are *stand alone*, in the sense that they are meant to be exercises for the student rather than work to be integrated with other activities in a classroom.

In this article, we describe a teaching system, DIAGNOSER, that both is based on a psychological/educational theory and provides an explicit interface into teacher-oriented activities. DIAGNOSER is meant to address three goals. First, we want to move assessment into the classroom in such a way that the tests and results are timely and useful to the teacher and student in judging the student's progress. Second, we attempt to tailor feedback to the student on the basis of his or her individual misconceptions. Third, we analyze the data resulting from real classroom use for patterns that tell us more about the concepts held by students, resulting in improvements to the system and, possibly, to psychological theory.

In order to understand the motivation for DIAGNOSER, it will help to look at standard methods of assessment. Traditionally, classroom assessment comes in two forms. There is some loosely structured and unstandardized day-to-day assessment by the teacher, which may take the form of paper quizzes and tests, conversational interactions, or other instances of graded performance. This type of assessment has the advantage of being immediate, in depth, and curriculum relevant. It measures what the teacher wants the students to learn, when the teacher wants to know about it. It is, however, labor intensive. It is limited by the teacher's expertise and the amount of time he or she has in contact with each student, as well as the amount of time he or she has to spend grading student work. When a teacher has over a 100 students to teach each day, this becomes a serious constraint.

Another difficulty with classroom quizzes is that performance on one teacher's assessments is often incomparable with performance in the class down the hall, much less a class in another school, district, or state. The teacher may attempt to "curve" grades to an expected distribution for the class or may make up items according to a rigid set of standards based on his or her interpretation of the curriculum, but both of these methods allow for substantial variation. Nor are these results generally available to the teacher down the hall to look at, much less researchers interested in student learning.

The other form of assessment is, to use Mislevy's term, "drop in from the sky" testing, in which students are temporarily isolated and monopolized for a very exact and standard measure. These standardized tests are commonly administered by the district or the state and are

This research was supported by the National Science Foundation, Grant REC9972999, to the University of Washington (E.H., principal investigator). Previous support for the development of DIAGNOSER was received from the James S. McDonnell Foundation. We thank Pamela Kraus, Aurora Graf, Dorin Popa, Virginia Bunch, Brianna Klink, and the participating teachers in the state of Washington for their assistance in this project. Correspondence concerning this article should be addressed to A. Thissen-Roe, Department of Psychology, University of Washington, Box 351525, Seattle, WA 98195-1525 (e-mail: thissen@u.washington.edu).

capable of measuring a student with great accuracy against the thousands or tens of thousands of students who take the test. (Although there are some national tests with larger samples, the state is usually the largest unit of standard academic administration.) They are written according to the curriculum as specified by the state, which means they are generally broad surveys rather than in-depth investigations of performance on a narrow topic. Thus, although they do not have the detail or the immediate relevance of the single-class assessments, they make solid general statements about a student's competence.

Another form of immediacy is lost entirely, however. Results from these tests often take months or even half a year to come back to the teachers. Tests given in the spring result in feedback to the teacher and student in the fall, when the student is in another grade. Although this kind of feedback may be useful to the teacher for calibration and to the school or district for tracking a student's lifetime progress, it does not allow the teacher to address particular problem areas or identify struggling students in time to help them (Hunt, 2002). Finally, although the teacher is not responsible for grading these exams, administration of them may take away days of class time.

By combining the advantages of both types of testing, we seek to develop a system of low-stakes, in-class quizzes that grade in a standard way across one state and that deliver immediate feedback to both the teacher and the student about how well the student is learning. The most efficient way to do this seems to be to have computers administer and grade the assessments.

Testing by computer is already common in many settings, such as "written" tests for driver's licenses or for admission to graduate and professional schools. In some cases, the computer simply reduces scoring and filing workloads and, thus, streamlines and speeds the procedure of testing. In other cases, the computer may adapt to the test taker, selecting items based on prior responses for a shorter testing time with equal accuracy. Finally, a computer can make some measurements that a paper-and-pencil exam cannot record, such as reaction time. It is best at scoring objective measures but may be able to substitute for some kinds of traditionally subjective measures, such as natural language short answer and essay grading, as technology improves (Carlson & Tanimoto, 2003; Graesser et al., 2004).

In order that learning time is not reduced when room is made for all this testing, the test itself can be made into a learning experience. Supplementary instruction is integrated with the test itself—a kind of "first aid" given for problems. Whether a student makes a correct or an incorrect response, the system must provide feedback that includes why the response was correct or incorrect at a conceptual level. That is, the feedback for an error is tailored to that particular error. At times, the test should integrate information from several related items to provide more complex feedback on general mastery of the topic.

This computerized instruction is not intended to replace interaction with a teacher but, rather, to complement it. It can be thought of as an immediate hint, a pointer in the right direction. Therefore, it is important that feedback be provided to both the student and the teacher. Furthermore, the feedback provided for the teacher should go beyond grading the student as having attained some level of mastery. The teacher needs information about the level of understanding of an individual student or a class, including a qualitative discussion of misunderstandings. All but the most experienced teachers may also need some indication of what they should do to deal with a particular misunderstanding. Both the teacher and the student will profit from a qualitative discussion of the student's understanding of a topic.

Adapting the feedback and subsequent items of the test to a student's performance is not a new concept. In the traditional conversational mode of assessment and teaching, teachers pick up the pace or slow down as needed and also may adjust their explanations when they notice a common misconception. For example, if a student learning about Newton's laws asserts that a sliding object stopped moving because it was not being pushed anymore, a teacher has alternatives to simply reiterating that acceleration is proportional to force applied. She might say, "Does a hockey puck stop as fast on ice as on a rug? There's something about the floor that's causing it to stop. That something is the force of friction . . ."

This type of interaction is the keystone of facet-based instruction, a teaching method developed by Minstrell (1992, 2001), in which student responses are believed to be diagnostic of underlying reasoning about narrow classes of situations but these classes are not dependent on surface features of the item. That is, students reason similarly across contexts on the basis of shared abstract features but may not have a completely consistent theory that explains every situation they might encounter. For some types of problems, students may sound more like Newton, whereas on another topic their reasoning may be more like that of Aristotle. This is similar to diSessa's idea of *p-prims* (diSessa, 1993).

In facet-based instruction, the teacher attempts to identify these narrow concepts or procedures for problem solving, both called *facets*, in each student. He or she then convinces the student to replace incorrect or problematic facets with facets closer to those a well-trained modern scientist holds. To do this, he or she may provide a counterexample or a situation the student's facet cannot explain or will produce a nonsensical answer to.

This technique is highly effective (Hunt & Minstrell, 1996). However, it is fairly demanding. In order to utilize facet-based instruction, a teacher must have a *catalog* of common misconceptions and good counterexamples. Although many experienced teachers have this sort of mental catalog, many do not. The problem is especially severe for newer teachers or those teaching outside their areas of expertise. Facet-based instruction, delivered solely by a teacher, is extremely time demanding. This is a serious problem since, as anyone only minimally familiar with the U.S. educational system can attest, time is precisely what teachers do not have.

By using an abbreviated form of facet-based instruction, a computer program can guide a student's learning toward more productive and canonically correct forms of reasoning about situations. Responses can be categorized into those indicative of different facets, rather than being marked as right or wrong. This is a partial solution to the problem of the many possible responses a student might make to an open-ended item (or wish to make to a multiple-choice item). Our aims here are in line with those of Graesser's AutoTutor project (described elsewhere in this issue). We admit that computer programs are, in general, less flexible than human teachers and, certainly, less flexible than an experienced human teacher. However, a computerized teaching system does not have to be infinitely flexible in order to be useful. All it has to do is to be able to deal with the responses that students make most of the time.

For that very reason, though, it is important that the system not be regarded as a fixed one. A good automatic teaching system must contain within itself information that can be used to improve its content. As in the case of the desirability of prompt feedback, this is hardly a new principle. Traditional testing programs, such as the College Board and various industrial tests, regularly gather item statistics and use them to improve testing. The same principle applies to testing associated with facet-based instruction. However, the problem of defining item statistics is more complicated than it is for traditional testing formats, for the issue is not simply whether a response is right or wrong, it is how consistently and in what contexts that response is encountered.

We now will turn to a technical discussion of the design of the DIAGNOSER program and outline how we have handled the problems just described.

System Specifications

The present DIAGNOSER is a modern descendant of an early DIAGNOSER program for delivering facet-based instruction (Levidow, Hunt, & McKee, 1991). The earlier version was intended for use with stand-alone computers, the appropriate technology for the time. The DIAGNOSER described here is a World-Wide Web program that is embedded technically within one system of programs and conceptually within a particular educational system. The ensemble of programs consists of the DIAGNOSER *proper*, which deals directly with the student, and a system for student registration, record keeping, and reporting and, conceptually most important, a set of guides to teachers. These guides are intended to place student use of DIAGNOSER in the context of other classroom instruction.

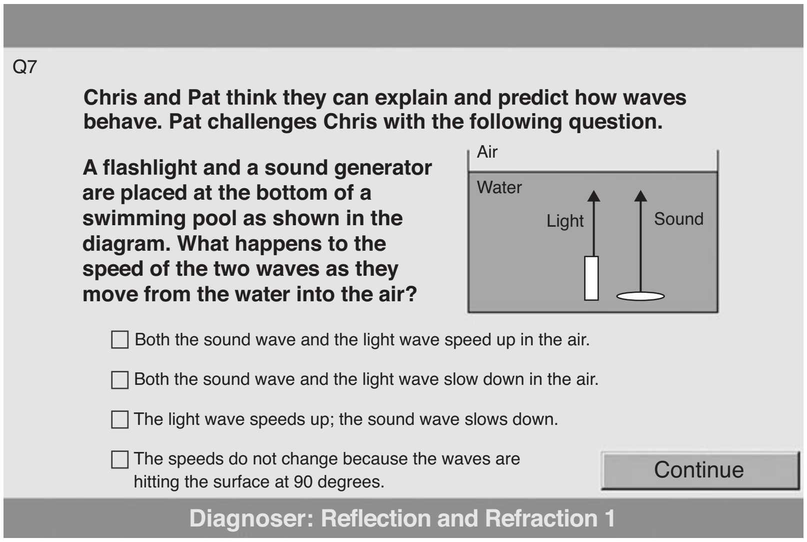
Conceptually, DIAGNOSER is embedded within the State of Washington standards for education, the state's "Essential Academic Learning Requirements" (EALRs). Embedding is accomplished by inserting into the teacher's guide pointers indicating which part of the DIAGNOSER content relates to statements in the state standards. Such information is extremely important to teachers, because the EALR standards are used to construct statewide examinations used to make policy decisions both about student progress and about the efficacy of instruction within a particular school or district.

A design principle that was followed throughout was that DIAGNOSER should be technologically simple enough that it could run on existing hardware in the schools. This stands in contrast to such systems as the Algebra Tutor (Milson, Lewis, & Anderson, 1990) and other ACT descendants, which are similar in spirit but count on introducing technology where it may not exist. Maintaining simplicity proved to be a daunting challenge and turned out to place severe limitations on the program, because of the great variety of hardware and software that we encountered as we made the system available throughout the state.

DIAGNOSER is Web-based and follows a client-server architecture with two client interfaces. The first Web interface is for students using the system to learn. This interface is relatively simple. Students using Netscape Navigator or Microsoft Internet Explorer can log in to personal accounts created for them by their teachers. Once in the system, they can see a list of question sets assigned to them and complete them one at a time. This part of the student–program interaction is maintained by a set of Active Server Pages.

Once a student selects a lesson, he or she is transferred to the DIAGNOSER itself. DIAGNOSER lessons (called *question sets*) were written in TOOLBOOK II,[1] a special program for delivering computer-presented instructions. Each question set appears in a separate window without browser controls and administers approximately 5–10 questions on a narrow topic, such as measuring and calculating acceleration or the interference of waves. Questions are usually multiple choice, with choices based on frequent student responses to similar problems posed on paper in an open-ended format. Some questions, however, require the student to enter a number without constraints, and a few ask the student to type out an explanation in complete sentences. Numeric responses are classified according to facet and allow a broad range of facets to be diagnosed, often more than will fit on the screen as multiple-choice options. Processing of free-text responses for facet content is not yet fully implemented in the system but has been automated for some items (Carlson & Tanimoto, 2003).

After a student solves a problem, the system either provides feedback according to the facet the student appears to hold or asks the student to explain the reasoning behind the answer. Usually, these second questions are multiple choice, even if the first question involved a numeric solution, and provide options that allow the student to confirm the system's diagnosis about the nature of his or her reasoning on that item. The answers provided for the second question are compared with the answer given to the first question. If the student chooses an explanation that does not lead to the answer he or she



Q7

Chris and Pat think they can explain and predict how waves behave. Pat challenges Chris with the following question.

A flashlight and a sound generator are placed at the bottom of a swimming pool as shown in the diagram. What happens to the speed of the two waves as they move from the water into the air?

□ Both the sound wave and the light wave speed up in the air.

□ Both the sound wave and the light wave slow down in the air.

□ The light wave speeds up; the sound wave slows down.

□ The speeds do not change because the waves are hitting the surface at 90 degrees.

Figure 1. An example of a physics problem in DIAGNOSER.

provided, he or she receives a warning from the system to be consistent. After such a question pair, some facet-based feedback is always provided.

The order of item presentation is neither completely adaptive nor completely fixed. All students begin with the same item. At some points in the question set, the student's path through the test branches, depending on the answers to previous questions. In many cases, this is to give students a chance to attempt an item a second time after feedback, although recording both attempts for the teacher and disallowing indefinitely repeated attempts. Another function of the branching is to provide different reasoning options to students who produced different answers to an initial question. Finally, the branching may either provide more in-depth diagnosis of a narrower content area for a struggling student or cut the quiz short for one who excels at the first several questions.

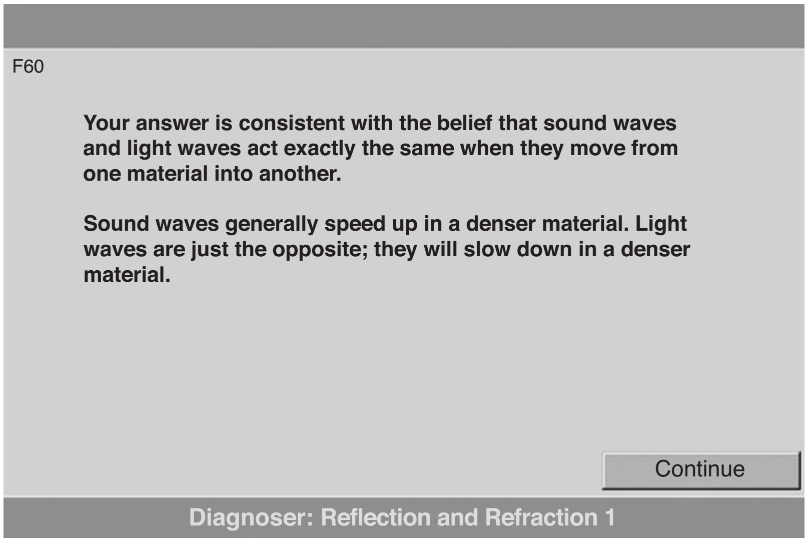
The items in DIAGNOSER were written by experienced high school physics teachers around the state of Washington, and the responses and branching patterns were chosen by the same teachers. This form of question generation was appropriate because, after all, the teachers were the experts who provided the knowledge for the expert systems program. However, this posed a problem. During development, we found that the teachers, although expert in teaching, were not familiar with programming concepts. As a result, they not infrequently provided branching patterns that were inconsistent or led to undefined paths. Therefore, a template was developed to specify branching patterns. A group leader who had participated in question writing and who was familiar with programming concepts filled out the template, indicating the branching pattern for a set. A programmer then transformed the template into a TOOLBOOK question set, using agreed-upon conventions for program writing.

We stress that creation of the question set is never complete. There must be a continual iterative process that, ideally, involves repeated revisions according to the patterns of observed data from a "live" system.

When a student has completed a set, his or her responses and the system's diagnoses are passed via HTTP POST to the DIAGNOSER server, where they are recorded in a Microsoft Access central database that manages accounts and assignments. This database will be described in more detail later. DIAGNOSER records the facets displayed by a student but does not attempt to assign grades in the conventional sense.

Teachers also log into the system by way of a Web browser. The teacher accounts, however, are kept separate and involve a more complex interface. In addition to trying out the question sets, creating and managing student accounts, and making assignments, teachers may review their students' performance immediately upon those students' completion of the question sets. Furthermore, teachers are provided with a reference called the *Teacher's Guide*. This is a set of Web pages that allow the teachers to learn about facets relevant to a topic currently under study (e.g., force and motion or the nature of gravity) or facet-based instruction in general. Suggested lessons are provided for activities both leading up to and following administration of the question sets. This way, a teacher who finds that he or she has a class full of students demonstrating a particular facet has a *prescription* at hand—that is, a lesson that addresses that misconception. The material in the *Teacher's Guide* was prepared after consultation with the same teachers who provided the question sets. The final construction of the guide was the responsibility of the third author (J.M.), who is himself an experienced high school physics teacher.

Although numerous details about each student's responses are recorded, including the individual choices of



F60

Your answer is consistent with the belief that sound waves and light waves act exactly the same when they move from one material into another.

Sound waves generally speed up in a denser material. Light waves are just the opposite; they will slow down in a denser material.

Figure 2. An example of feedback given to a student by DIAGNOSER.

a response to multiple-choice questions and any free-form comments, Active Server Pages allow us to provide a more concise report when a teacher requests information about student performance. The teacher sees a table of facets diagnosed by the system, within each set, with a row for each student and a column for each item. Each facet is represented by a brief number code that is easy to recognize. A hyperlink allows the teachers to see the text definitions that correspond to the numbered facets. That way, a teacher can note that his or her class is displaying Facet 40 frequently, particularly on Item 5, and look back to see both what Facet 40 was and what Item 5 said to elicit that facet.

DIAGNOSER and all its associated Web pages are served off a single PC with a fast network connection. The DIAGNOSER server runs Microsoft Windows, with Microsoft Internet Information Server delivering the content and Microsoft Access managing the database. Although a more robust or heavy-duty system could certainly be found, performance testing and active use by a large number of students have shown that the existing simple system is “good enough.”

DIAGNOSER question sets are delivered using a combination of HTML, Java, and JavaScript, using primarily text, simple widgets, and low-bandwidth images. Although the server has a relatively large amount of bandwidth available to it, the client computers often do not. Some students may access DIAGNOSER sets as homework by way of dial-up connections; some entire school districts are limited to a single slow connection to the outside. It is important to make the sets small enough in data size that they respond quickly even under these conditions.

In addition to providing the teachers’ guide, the DIAGNOSER project is supported by a Help telephone line. Help calls fall into three broad categories: requests for minor assistance in using DIAGNOSER (e.g., a teacher has forgotten his or her password), requests for assistance concerning instruction (which we refer to experienced teachers working with us), or requests for more serious technical assistance. Some of the latter calls reflect the varied nature of hardware and software support in schools. For instance, although some districts have excellent technical support, others do not. We have received calls from teachers who were using Web browsers that were years out of date and who did not know that free upgrades were available.

Some technical support calls are more serious. Most of these have to do with interactions between the program and the server in the teacher’s school district. The Web delivery of item content and the return of response data via HTTP POST keep the entire data transaction on Port 80, allowing DIAGNOSER to operate across many firewalls that restrict specialized Internet information exchanges. Content filters still are responsible for a number of our technical support calls; well-meaning districts or schools may choose to limit student access to sites that have met with prior approval, in the interest of protecting the students from content of which their parents may disapprove. In this case, the DIAGNOSER team works with information technology personnel in the district or school to get permission instituted for the site.

The extensive records of the student response database are also immediately and usefully available to the research team. In addition to traditional item response parameters, interitem consistency can be measured. That is, if a student demonstrates the same facet across several questions, he or she is responding consistently. Consistency of this type suggests a well-defined facet that is meaningful and stable enough to be addressed by counterexample feedback and additional lessons. A lack of consistency can indicate that a facet is poorly defined, being either too narrow or too broad in scope, or that it is not stable. That is, students may be learning between opportunities to display the facet. Learning during the administration of the question set is to be desired, but it also means that the student's concepts are a moving target and more difficult for the teacher to address correctly.

In order to deal with analysis of consistency and learning, we have had to develop some new psychometric approaches to data analysis. These will be reported in a subsequent article.

Another statistic easily gathered from the database is overall frequency of selection of each multiple-choice response and frequency of diagnosis of numeric responses. If many students are choosing *none of the above* type responses, or producing numeric responses that cannot be clearly diagnosed, an item is not providing much information. Therefore, consistency and response frequency make up important criteria for the revision process.

Usage and Validation

DIAGNOSER is in its 3rd year of active use. An estimated 6,000 students used the system during the 2002–2003 school year, up from 1,080 during the 2001– 2002 pilot year. It may seem strange that a computer system would record only approximations regarding use, so we shall explain why the exact number is difficult to determine. When a teacher registers a student, an account is created for that student, and of course we know how many accounts have been created. However, the number of student accounts created is more than the number of students actively using the system, because many teachers create "extra" accounts either to explore the system (which we encourage) or because they intend to assign a lesson and then decide not to do so. On the other hand, the number of accounts with a corresponding completed question set is lower than the number of students using the system, because many, if not most, teachers using the system instruct their students to work in pairs or groups. Indeed, the teachers who worked with us in creating DIAGNOSER often do this themselves, in order to benefit from cooperative work.

Initial results regarding the frequency of display of facets were surprising to teachers, even those who constructed the sets. We found that teachers tend to overestimate the frequency of rare misconceptions, while underestimating the frequency of common ones, particularly those that come up nearly as often as the correct answer. Revisions to some question sets are underway, incorporating these data as well as data about how teachers use the sets.

A study of consistency of responding in the 2001–2002 and 2002–2003 databases revealed both consistency and inconsistency. Some facets and some sets performed much better than others. On some topics, only one or two facets beyond the correct, or "expert," facet differentiated themselves from the mass of wrong answers, suggesting that the students may have been guessing or responding on the basis of surface features only. On other topics, a distance measure of inconsistency showed facets to occupy separate conceptual spaces. Students, therefore, appeared to consistently prefer one wrong procedure or concept to other distractors (or had mastered the subject entirely).

One Washington state school district participated in a validation study. The results of this study will be presented elsewhere. Briefly, the state of Washington's annual examination contained several questions on topics covered by DIAGNOSER question sets. Students who used DIAGNOSER on multiple occasions outperformed their peers by 14 percentile points, a substantial gain. These validation results, although promising, cannot discriminate between two possible explanations. The system may have assisted the students in learning and their teachers in teaching. Alternatively, teachers who are better to begin with may have chosen DIAGNOSER for use in their classrooms. Further research is needed, but either way, we are convinced of the need to offer such a system.

Conclusions

In summary, we have developed a system that fits our initial goals. We provide timely topic-relevant in-class assessment by way of computer-administered quizzes. The results of these assessments are both immediate and standard across classrooms. Facet-based instruction allows us to provide tailored feedback according to student misconceptions. Finally, in the process of administering this service, we accumulate a database of student responses. Patterns found in this database, such as consistency and facet frequency, can inform future iterations of the system and our theories of concept development.

REFERENCES

Carlson, A., & Tanimoto, S. L. (2003). Learning to identify student preconceptions from text. In J. Burstein & C. Leacock (Eds.), *HLT-NAACL 2003 Workshop: Building Educational Applications Using Natural Language Processing* (pp. 9-16). Edmonton: Association for Computational Linguistics.

diSessa, A. A. (1993). Toward an epistemology of physics. *Cognition & Instruction*, **10**, 105-225.

Graesser, A. C., Lu, S., Jackson, G. T., Mitchell, H. H., Ventura, M., Olney, A., & Louwerse, M. M. (2004). AutoTutor: A tutor with dialogue in natural language. *Behavior Research Methods, Instruments, & Computers*, **36**, 180-192.

Hunt, E. (2002). The DIAGNOSER project: Assessment in the service of learning. *Journal of the Hellenic Psychological Society*, **9**, 241-251.

Hunt, E., & Minstrell, J. (1996). Effective instruction in science and mathematics: Psychological principles and social constraints. *Issues in Education: Contributions from Educational Psychology*, **2**, 123-162.

Levidow, B. B., Hunt, E., & McKee, C. (1991). The DIAGNOSER: A HyperCard tool for building theoretically based tutorials. *Behavior Research Methods, Instruments, & Computers*, **23**, 249-252.

Milson, R., Lewis, M. W., & Anderson, J. R. (1990). The teacher's apprentice project: Building an algebra tutor. In R. Freedle (Ed.), *Artificial intelligence and the future of testing* (pp. 53-71). Hillsdale, NJ: Erlbaum.

Minstrell, J. (1992). Facets of students' knowledge and relevant instruction. In R. Druit, F. Goldberg, & H. Niedderer (Eds.), *Research in physics learning: Theoretical issues and empirical studies* (pp. 110-128). Kiel: University of Kiel, Institute for Science Education.

Minstrell, J. (2001). The role of the teacher in making sense of classroom experience and effecting better learning. In S. M. Carver & D. Klahr (Eds.), *Cognition and instruction: Twenty-five years of progress* (pp. 121-150). Mahwah, NJ: Erlbaum.

NOTE

1. Toolbook is a product developed by Click-To-Learn, Inc. (formerly, Asymmetrix Inc.).

(Manuscript received November 15, 2003;
accepted for publication March 10, 2004.)