The DIAGNOSER project: Assessment in the service of learning

EARL HUNT
University of Washington, U.S.A.

ABSTRACT Formative assessment is intended to aid student learning, rather than to evaluate the student for purpose of selection or prediction of performance. The DIAGNOSER project has designed formative assessments using computer-presentations. The idea behind the DIAGNOSER, FACET-BASED INSTRUCTION, is that students come to a topic with pre-formed ideas. The goal of assessment is to identify these ideas and provide feedback tailored to the student's current ideas, rather than just responding by telling the student that the answer was right or wrong. This method of assessment represents the student as being at a location in a space of knowledge states, rather than representing the student by a vector of factor scores. DIAGNOSER modules have been developed for topics in the physical sciences and in statistics, and have proven quite successful. DIAGNOSER modules are now being placed on the World Wide Web and are being coordinated with a program of high-stakes testing.

Key words: Assessment, DIAGNOSER, Facet-based instruction.

Introduction

The topic I want to talk about is the assessment of mental competence, and particularly the assessment of competence in science and mathematics. This sort of assessment is traditionally seen as an educational endeavor. My discussion will introduce topics from psychology and computer science as well as education. I hope that the methods described here will point the way toward a rather different view of assessment than the classic one.

To explain why I feel this way we first have to look at what traditional assessment is. In a presentation to the U.S. National Research Council's committee on cognitive assessment Robert Mislevy referred to the traditional technique as "Drop in from the sky" assessment. Assessors from the Education Ministry, clinical psychologists, or whatever suddenly appear in the person's life, sit them down for anywhere from a few minutes to a day, ask some questions that seem to come out

---

Note 1. These are the speaking notes for a presentation to the 5th European Conference on Assessment (Patras, Greece: August 1999). The DIAGNOSER project has been supported by the James S. McDonnell Foundation and the National Science Foundation.

Note 2. DIAGNOSER is a technological development that grew out of research on facet-based instruction, an educational method pioneered by Dr. Jim Minstrell. I am more than happy to acknowledge his contributions to my thinking on this project, and about education in general. I also wish to acknowledge the assistance, advice, and collegiality of David Madigan, Bjorn Levidow, Andrew Schaffner, Aurora Graf, Jessica Baldis, and the many teachers who have contributed to the DIAGNOSER projects.

Address: Earl Hunt, Department of Psychology, University of Washington, Box 351525, Seattle, WA 98195-1525, U.S.A. E-mail: ehunt@u.washington.edu

of nowhere, and then go away. Some time later the person being assessed finds out that the evaluator scored the answers to the questions, and issued a report. The examinee may even receive that report, although it is more likely that he or she will simply receive a score.

In spite of my (and Mislevy's) somewhat pejorative term, 'drop in from the sky' assessment has its purposes. It is useful in personnel selection, for it provides a cost-effective way of screening applicants. (This application, of course, includes selection for educational advancement.) Assessment independent of further interaction with the examinee is also an effective way to certify individual accomplishment. There are good arguments for requiring that things like the U.S. Bar Examination and Medical Board certifications be conducted by agencies who had nothing to do with training the people to be certified. Finally, and increasingly, assessment is useful for the purposes of program evaluation. If the students from a particular school repeatedly fail to meet reasonable assessment criteria the school's governors should ask some pointed questions of the faculty, and perhaps of themselves. From the viewpoint of educational and training institutions, assessment is a form of quality control.

The idea that assessment is part of quality control applies to the individual as well as an institution. In educational and training situations assessment ought to be a tool in the service of learning. It ought to go beyond telling someone that s/he does or does not know something, it ought to assist the examinee in furthering his/her competencies. In order to attain this goal assessment has to be connected to useful and timely feedback. Here I introduce my first psychological principle. For informative feedback to be useful it has to be almost immediate. Why? Contrary to what clinical psychologists will tell you, negative feedback is a good thing. Or at least, negative feedback in the engineering sense (where the term was first introduced) is a good thing. Negative feedback is a signal that compares the desired response with the appropriate response, in order to illustrate the difference between them. Negative feedback is useful to the extent that the individual can use it to inform him/herself about how responses should be computed in this and similar situations.

This brings me to my second psychological principle. When people respond to difficult questions, as they have to in scientific and mathematical problem solving, they compute the answer based upon their understanding of the principles involved. We can represent these understandings as schemas. The existence of correct reasoning schemas in scientific reasoning has been documented over and over again. Wrong answers are seldom simply random deviations from the correct schema. They represent semi-orderly thought that is not quite what Newton, Einstein, and the mathematics teacher had in mind. In order for educational and training assessment to be useful in the service of learning the assessment has to identify the schemas that the student is using and provide feedback that modify these schemas to be more in accord with currently-held views of 'correct' scientific and mathematical reasoning.

These remarks suggest a sort of New Age view, that all knowledge is provisional. While this is true, in a sense, in science some ideas are less provisional than other. Newtonian mechanics, Einstein's theory of relativity, and Darwin's theory of evolution do not have the same provisional status as the latest theory in the social sciences. In mathematics, of course, truth is not provisional at all. Assessment should bring people closer to our best understanding. And simply telling them what the right answer is (which is what didactic instruction does) seldom works. The educated person knows why the 'right' answer is more correct than the alternatives. That is the sort of person that education should produce. Properly organized assessment can help.

Facet-based Instruction

As my opening remarks implied, we are con-

cerned with assessment in the aid of learning. This means that we have to have a model of the learning process. The model that is behind our assessment research is facet-based instruction, an educational model pioneered by my colleague Jim Minstrell (Hunt & Minstrell, 1994; Minstrell, 1992). The idea is based on two assumptions:

When students approach a topic in science and mathematics (and I think virtually everywhere) they bring with them a set of beliefs about how problems should be solved in the content area under discussion. Following Minstrell, I will refer to these ideas as facets. Facets are not well worked out, but erroneous theories (as Aristotelean physics was.) Rather, they are somewhat limited notions of how to deal with certain situations. A good example is the idea that energy expenditure and power, in a non-technical sense, are associated with force production. Note that this makes 'force' a property of the agent, which is not in accord with the physicist's definition of force, as a relationship between two objects. The 'force production' notion of reasoning makes it hard to realize that when a man runs into a truck, the man and the truck exert equal and opposite forces on each other. Note, though, that the force as energy production notion is not totally wrong. And pragmatically, it is useful. If you act in accordance with this rule you will avoid running into trucks.

People do not learn (much) by being told the right answer. The good instructor does not recite right answers, or even offer proofs that they are right. The good instructor identifies the facets that the student applies to a particular situation. Then the instructor offers educational experiences that will convince the student to modify his or her current facets, hopefully in the directions of the facets that constitute our current understanding of physics and mathematics (for which we have data), or any other topic under discussion.

At this point I want to make a brief aside. I have been using the word 'student.' Students are not necessarily enrolled in a public school. The approach and methods that I describe here have been applied in U.S. middle school and high school science, university training in statistics, and even to assist in medical problem solving by licensed practitioners. My argument will be illustrated by school assessments but the method is more general.

Facet-based instruction faces the instructor with considerable challenges. The biggest one is that the teacher has to know more than the right answer. The teacher has to know what limited facets students may have, how to recognize them, and what to do when a student uses a problematical facet. Negative feedback is not equivalent to beating students about the shoulders, metaphorically or otherwise!

Many instructors simply do not have the information or training required to present facet-based instruction. This is particularly true of instructors who have limited experience and/or of instructors who deal with very large classes. (There may be a message here for advocates of distant learning!) Traditional instruction emphasizes good, well organized, didactic presentation of content, not engaging in an interaction with individual students. At least in the United States, and I think in many other countries, evaluation of teachers (including university instructors) is largely based on how up-to-date they are in their knowledge of the subject matter, and how well-organized they are in presenting it. Facet-based instruction requires a great deal more content-knowledge than traditional instruction does, because the teacher does not always get to set the context of a question. It places much less emphasis on well-organized delivery.

There are situations in which some teachers use traditional instructional techniques, even though they believe in and are qualified to do facet-based instruction. The reason is simple: time. An instructor with a class of 200-500 students, meeting three times a week, does not have time to engage in discussion with individuals. Obviously learning can take place in such situations;

universities work. But do they work as well as they should? And when we rely entirely on didactic instruction are we not transferring the cost of education? This is the point at which computer-based assessment in the service of learning can help.

This brings us to a program, the DIAGNOSER, which combines the assessment of student facets with instruction directed toward those facets. I stress as strongly as possible that the DIAGNOSER is not an artificial intelligence device intended to replace the teacher. (Indeed, I am very skeptical of proposals for this sort of education.) The DIAGNOSER is an expert-system program, where the expertise is in teaching rather than in knowledge of the topic. It is intended for use in conjuction with at least an approximation to facet-based instruction in the classroom.

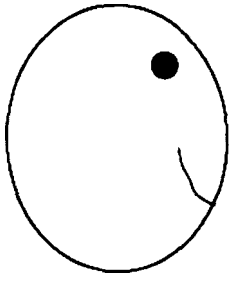
The organization of the DIAGNOSER

The student's view

Figure 1 shows the DIAGNOSER organization from a student's view. The program consists of a series of modules organized around some coherent topic, such as Kinematics, Nature of Gravity, Elementary Probability Theory, or The Water Cycle. The single question in a conventional test is replaced by a four-part sequence. The first part in the sequence is a phenomenological question. A typical one, taken from our Nature of Gravity module, is shown in Figure 2. Note that the question is in multiple choice form. The right answer, of course, is dictated by the science or mathematics involved. The wrong answers are more interesting, for they require psychology to define them.

Each wrong answer is keyed to a more-or-less problematical facet. In the Nature of Gravity two of the answers reflect the common beliefs that (a) the weight of a body depends only on gravity, regardless of the medium in which the body is weighed or (b) air presses down upon an object, so its absence makes a body lighter.

After the student has entered an answer to the phenomenological question, and before commenting further, the program asks the student a reasoning question. This is shown in Figure 3. The reasoning question is supposed to accomplish two goals. First, the fact that a reasoning question always follows a phenomenological question helps discourage a common tendency



Phenomenological question:

What will happen in the following situation?

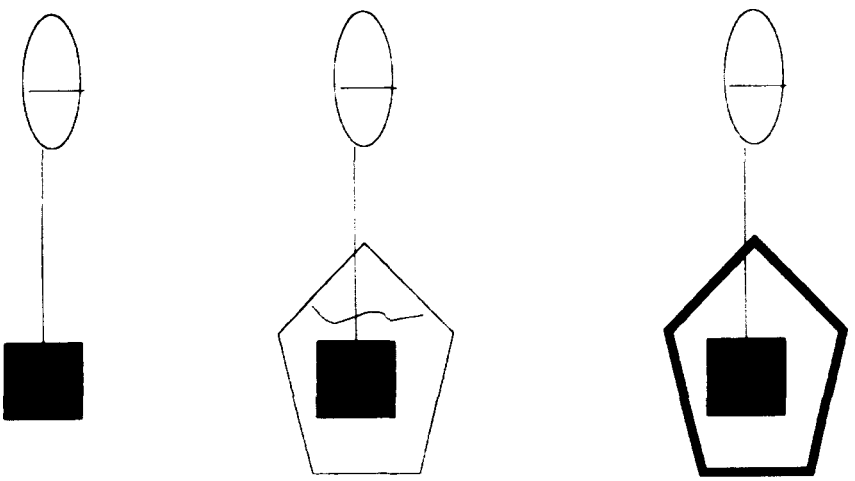
Reasoning question:

Why do you think that?

Diagnosis: Comment on student understanding.

Prescription: You might consider the following situations...

Figure 1
DIAGNOSER as it appears to the student.



A block of material that weighs 20 kg. when suspended in air (a) is also weighed when it is fully immersed in water (b) and in a vacuum (c). The scale is extremely accurate. Compared to the reading in air will the scale reading be

○ (a) The same in all three cases?

● (b) Less than 20 kg. when the object is in water and more than 20 kg. when the object is weighed in a vacuum?

○ (c) Less than 20 kg. when the object is in water and less than 20 kg. when the object is weighed in a vacuum.

Click here if you wish to comment:

Figure 2
An example of the first (phenomenological) question in a DIAGNOSER sequence. A possible student answer is marked with an x.

Which of the following reasons best captures your belief about the weights in the previous question?

○ (a) Weight is determined solely by gravity.

● (b) Air pressure pushes down on an object, so when it is removed the object is lighter.

○ (c) The water exerts a buoyant force upward, so the object weighs less in water. It does not matter whether you weigh the object in air or a vacuum.

○ (d) Both water and air exert a buoyant force upward, but the force exerted by water is much greater than the force exerted by the vacuum.

Click here if you wish to comment:

Figure 3
An example of the second (reasoning) question in a DIAGNOSER sequence. A possible student answer is marked with an x.

Your answers to the first two questions are consistent. However your reasoning does not work in many situations. Move ahead to the next screen for further comment.

Figure 4
An example of a reasoning panel that might be displayed if a student answered in the way that is shown in Figures 2 and 3.

to answer on the basis of 'intuitive' feelings. At least in science and mathematics, we want to encourage tightly controlled reasoning. Second, the answers to the reasoning questions give us (the program developers) a further bit of insight into student reasoning.

Note the comment section at the end of both the phenomenological and reasoning screens. If the student does not like any of the answer choice offered he or she can simply say so. Among other things, these comments provide us, as program developers, with insights into new facets.

The program then begins to comment. The "diagnosis" message, which always follows the reasoning question, first lets the student know whether the answers to the phenomenological and reasoning question were consistent with each other. In addition, we comment on whether or not the reasoning was, in fact, correct. Figure 4 shows the diagnosis if a student answered in the way shown by the square marks in Figures 2 and 3. We believe this is important, because science and mathematics reward consistent reasoning, and we want students to understand this. (As a passing comment on Psychology, it has been observed that one of the major problems

Don't things weigh less when they are in water than in the air? That is because the water exerts a buoyant force upwards, because the water pressure is greater at the lowest part of the object than at the highest (the point nearest the surface). The same thing happens with any other medium. For instance, the atmosphere can be thought of as a vast sea of air that surrounds the earth. Because air is not very dense it exerts a very slight upward force, but it is there!

Figure 5
An example of a prescription panel that might be displayed if a student answered in the way that is shown in Figures 2 and 3.

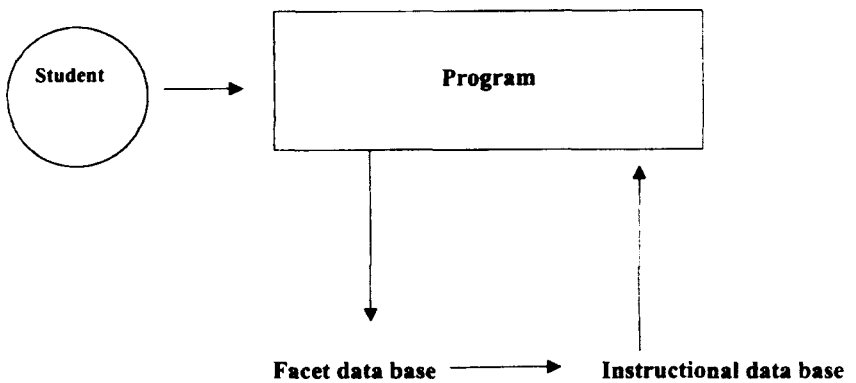


Figure 6
The program developer's view of DIAGNOSER.

people have in scientific reasoning is that they are not sensitive to the need for consistency between theory and observation [Kosłowski, 1996].)

Finally, the program provides a prescription *screen or screens*. The prescription screen is intended to move the student toward more correct reasoning, by showing how the student's original reasoning would lead to trouble, and then moving the student toward a more correct understanding. An example is shown in Figure 5.

We design modules so that they take from 10 to 30 minutes to answer. The reason for this is that teachers use the DIAGNOSER very much the way that physicians or dentists use technical assistants; to conduct an initial interview with the client. The program then summarizes the results, so that the teacher can read them. The report does not just say how many or which questions a student got right or wrong. The report identifies any problematical facets of reasoning that the student has displayed. Teachers find this useful in guiding their own further comments to the students. In addition, the report summarizes any comments that the student has made during the interaction.

The developer's view

Figure 6 shows what the DIAGNOSER looks like from the viewpoint of a program developer.

Student responses are treated as inquiries into a database containing common facets of student understanding. The response to these inquiries is determined first by accessing the appropriate prescription and diagnosis pages. In addition, a summarizing program keeps track of student interactions in order to prepare the teacher's report. This is delivered at the end of the DIAGNOSER session.

The power of the DIAGNOSER lies in the database of facets and the accuracy with which the questions draw forth particular facets.

Three sources are used to develop facets. First, whenever we can, we conduct extensive interviews with teachers. In order for this to be fruitful we have to contact a particular type of teacher; one who has the time and energy to listen carefully to students' ideas. Not all good teachers do this. As I pointed out earlier, it is possible to be an excellent didactic lecturer without listening to student ideas. Teachers who have listened to their students are often gold mines of information.

A second way that we find facets is by asking students to write out their own understandings of various phenomena. We do this prior to instruction, for we want to know what sort of ideas students have as they enter a class. Figure 7 shows one such example, used to develop the WATER DIAGNOSER. This method is less labor intensive than the first, because one or two people oriented toward facet-based instruction can read the students' answers. While it is not possible to give a definitive answer, reading on the order of 100 student responses seems to give us enough information to develop a useful facet list.

The third way of finding facets is simply to observe how students use the DIAGNOSER, what comments they make, and what questions they direct to teachers.

After having developed the facet list we have to develop prescriptions. This is the point at which teacher input is crucial. What we attempt to do in these situations is to identify a group of a half-dozen teachers who will work with us over a period of about a year in identifying facets and prescriptions. Since teachers are very busy people, especially in U.S. schools, finding skilled teachers to write prescriptions may be the weakest link in the entire approach.

We have explored developing prescriptions through the use of the psychological literature on analogical reasoning. Unfortunately, we find that the psychological literature is often deficient, for many of the experiments in it seem to be designed to demonstrate a theory rather than to address problems that occur in practical situations. This is quite understandable, the development of cognitive science depends upon the development of theory. However, I think there is room for the development of cognitive engineering, and engineering has to be more closely allied to practice than science is.

Results

DIAGNOSER type programs have been written for high school physics, middle school mathematics (dealing with real numbers, ratios, and proportions), university-level statistics, and middle school Earth Sciences.

You have a small cup of water and a large bag of sugar. You begin mixing sugar into the water. At first the suggar appears to disappear and the water remains clear. After you have mixed several tablespoons of sugar into the cup of water, however, the sugar no longer disappears, but apprears to stay suspended in the water. Explain what is happening and why.

Figure 7
An example eliciting questions.

The usual use of the program is for self-assessment by the student and as an advisory to teachers. One of the best uses I have seen was in a situation where the instructor split the class into groups, each around a computer, and had the students go through the program as a group. This facilitated a great deal of discussion between the students, which was what was intended. When students got through with the questions in a content module they called the teacher over, the teacher looked at the summary (much as a physician might look at the report from a medical laboratory) and selected his/her instruction appropriately. Several studies have shown that when the teacher uses facet-based instruction, combined with DIAGNOSER, students perform considerably better than they do in comparable control classes (Hunt & Minstrell, 1994, 1996). These comparisons mix the effect of the teacher and classroom interactions, which we think are probably the most important aspect of the program, with the effects that are due solely to the use of the DIAGNOSER program.

We have looked at a few situations in which the DIAGNOSER program was used in isolation. We have been able to obtain improvements in both University level statistics classes and a Middle-School science course when DIAGNOSER is used as a review program. These represent the high and low ends of our applications. In both cases the program produced better learning than an appropriate control for instruction and/or review. These are illustrated by the Middle School results shown in Figure 8. The university results were comparable, but the analysis was too detailed to go into here.

Discussion

Like good (cognitive) engineers, we believe that we have now established "proof of concept."

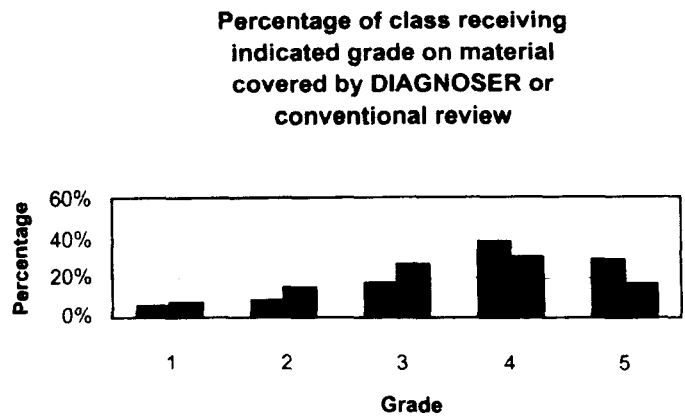


Figure 8
Results on a test on a Middle-School science topic (water cycles) after review using DIAGNOSER or conventional techniques.

Note: The relative frequency of answers for DIAGNOSER classes are shown in dark bars, for comparable controls in light bars. Grades assigned were from 1 (best) to 5 (worst). Grade 5 was considered unsatisfactory by the school district involved. Grade 1 indicates better than 90% correct, grade 2 from 80 to 90% correct, and so forth.

Our next step is to go into full production. We are now embarked in a program that uses the DIAGNOSER throughout an entire educational system.

The State of Washington (U.S.A.) has developed a set of educational standards for public schools. All students in the state are assessed once in primary school, once at the end of middle school, and once in the 10th grade. The results of these tests are intended to be used for student qualification and program assessment, so they are classic "Drop in from the sky" tests. The results have a great deal of importance for students, teachers, and school administrators.

In conjunction with the state's Office of Public Instruction, we are developing world wide web (www) based versions of the DIAGNOSER that are intended as self-assessment guides, in the service of learning for both students and teachers. The material covered will be mathematics and physical sciences as specified by state standards for the middle and high school level. The logic of the program will be essentially the logic I have described, except that the rigid question-reasoning-diagnosis-prescription sequence will be relaxed to provide for multiple screen presentation at each point. Therefore what we are developing would be better called scenarios than questions. In addition to providing prescriptions for individuals, the program will provide suggested class exercises for teachers, keyed to the facets that have been displayed by the students in the teacher's class.

Participation in this program will be voluntary. The program as a whole will be evaluated by the performance of students on the state assessment examinations. Since we will be dealing with a very large statistical base, it will be possible to institute elaborate statistical controls to evaluate such influences as teacher experience, composition of the student body, and so forth. The project can be considered an attempt to apply cognitive engineering, centered around assessment in the service of learning, in a very large setting.

We hope to have trial programs running in about 9-10 months. At that time anyone will be able to log on and look at what we are doing.

Psychometric issues

Since the paper was originally presented in a conference on assessment, I should like to close with a challenge to psychometricians. The standard model of assessment can be thought of as trying to locate a person in a "mental space" defined by appropriate dimensions of cognition or personality. This remark applies to the Gc-Gf model of intelligence, the "Big Five" model of personality, and many other psychometrically oriented theories. There is an implicit assumption in the related assessment efforts that the testing process does not itself change the intellectual or personality status of the person being tested.

The DIAGNOSER project rejects both of these views. First, we do not think of a person as being located in a space. We think of a person as being in a particular belief state. The correct mathematical analogy would be to nodes in a network. We want to find out what node represents the current belief state, find the shortest path to the node representing the desired belief state, and then do what we can to move the person toward that path.

The problem of isolating a person's belief state is not intractable, in theory. Bayesian analytic techniques have been proposed as a way to solve the problem. However, these techniques typically require a lot of testing in order to rule out different hypotheses. The testing itself, being in the service of learning, is likely to move a person from one node to another. At least, we hope it does! But this means that the psychometric estimates are shooting at a moving target.

I regard this as a challenge for psychometrics, rather than a reason to abandon the DIAGNOSER approach.

Conclusion

The DIAGNOSER project combines the Artificial Intelligence concept of expert systems with psychological ideas about schematic and analogi-

cal reasoning. We then attempt to put the whole thing into an educational context. I cannot say that the scientific basis tightly constrains our work, for most of cognitive psychology is not sufficiently precise to do that. While we use cognitive psychology as a guide to action, we combine theory with a long and equally challenging project of capturing the knowledge of experienced teachers. That is all right with us. Lesgold and Nahemow (2001) has remarked that education and training are really cognitive engineering. We agree, and we hope that we are building a better (educational) mousetrap. The challenge is formidable. If we succeed, I think that our view of belief states and movement between them may prove as influential on psychometrics in the 21st century as the factor analytic model has been in the 20th.

References

Hunt, E., & Minstrell, J. (1994). A cognitive approach to the teaching of Physics. In K. McGilly (Ed.), *Classroom lessons: Integrating cognitive theory and classroom practice* (51-74). Cambridge, MA: MIT Press.

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

Koslowski, B. (1996). *Theory and evidence: The development of scientific reasoning*. Cambridge, MA: MIT Press.

Lesgold, A. M., & Nahemow, M. (2001). Tools to assist learning by doing. In D. Klahr & S. Carver (Eds.), *Cognition and instruction: Twenty five years of progress* (pp. 307-346). Mahwah, NJ: Erlbaum.

Minstrell, J. (1992). Facets of students' knowledge and relevant instruction. In R. Duit, F. Goldberg, & H. Neidderer (Eds.), *Proceedings of an international workshop. Research in Physics Learning: Theoretical issues and empirical studies* (pp. 110-128). Kiel, Germany: Institute for Science Education.