COGNITIVE SCIENCE **18,** 439–477 (1994)

Eliciting Self-Explanations Improves Understanding

MICHELENE T. H. CHI, NICHOLAS DE LEEUW,
MEI-HUNG CHIU, AND CHRISTIAN LAVANCHER
University of Pittsburgh

Learning involves the integration of new information into existing knowledge. Generating explanations to oneself (self-explaining) facilitates that integration process. Previously, self-explanation has been shown to improve the acquisition of problem-solving skills when studying worked-out examples. This study extends that finding, showing that self-explanation can also be facilitative when it is explicitly promoted, in the context of learning declarative knowledge from an expository text. Without any extensive training, 14 eighth-grade students were merely asked to self-explain after reading each line of a passage on the human circulatory system. Ten students in the control group read the same text twice, but were not prompted to self-explain. All of the students were tested for their circulatory system knowledge before and after reading the text. The prompted group had a greater gain from the pretest to the posttest. Moreover, prompted students who generated a large number of self-explanations (the high explainers) learned with greater understanding than low explainers. Understanding was assessed by answering very complex questions and inducing the function of a component when it was only implicitly stated. Understanding was further captured by a mental model analysis of the self-explanation protocols. High explainers all achieved the correct mental model of the circulatory system, whereas many of the unprompted students as well as the low explainers did not. Three processing characteristics of self-explaining are considered as reasons for the gains in deeper understanding.

Preparation of this article was supported in part by an Office of Educational Research and Improvement Institutional grant for the Center of the Study of Learning. The opinions expressed do not necessarily reflect the position of the sponsoring agencies, and no official endorsement should be inferred.

We are grateful for comments provided by Kurt VanLehn, and by the action editor and reviewers, Keith Holyoak, Miriam Bassok, Allan Collins, and Pete Pirolli.

Correspondence and requests for repirnts should be sent to Michelene T. H. Chi, 821 Learning Research and Development Center, University of Pittsburgh, 3939 O'Hara Street, Pittsburgh, PA 15260. Requests by e-mail may be sent to chi@vms.cis.pitt.edu.

Mei-Hung Chiu is now an associate professor at the Graduate Institute of Science Education, National Taiwan Normal University, 88, Sec. 5 Roosevelt Road, Taipei, Taiwan 117.

Processes of learning are often considered in terms of either comprehension, skill acquisition, or both. In the case of comprehension, the declarative information that is to be understood (such as a story) maps onto knowledge that is already stored and organized in memory. The emphasis in comprehension is therefore on the instantiation of existing knowledge. In the case of learning a procedural skill, two classes of mechanisms are proposed: knowledge acquisition and compilation. In knowledge acquisition, an initial version of a skill (such as a set of instructions on how to send e-mail) is directly encoded from the source of instruction. In knowledge compilation, the encoded skill is slowly transformed so that it becomes more efficient. The emphasis in skill acquisition is, therefore, on the progression of a skill once encoded.

However, new instruction of either declarative or procedural knowledge cannot always be either instantiated or directly encoded; often it requires the *integration* of new information with existing knowledge. This integration process can be facilitated by asking students to actively construct what they are learning. In Chi, Bassok, Lewis, Reimann, and Glaser (1989), a special form of construction activity called "self-explaining" was shown to be effective at improving the acquisition of problem-solving skills. Apparently, spontaneously generating explanations to oneself as one studies worked-out examples from a text is a process that promotes skill acquisition even though it is neither the direct encoding of instruction nor the compilation of an encoded skill.

The original Chi et al. (1989) study concerned the learning of a procedural skill from examples provided in a physics text. Generally, a worked-out solution example takes the form of a *sequence of action statements*, such as "consider the knot...to be the body" or "choosing the x- and y-axes as shown" without any explanations or justifications for the actions chosen. It was hypothesized, therefore, that in order for students to learn maximally from sparsely stated worked-out examples, they must provide their own explanations for the actions in the examples. Eight college students, who first studied the prose sections of an introductory physics text were then asked to explain (on a voluntary basis) whatever they understood from reading statements from three examples. The basic result is termed the self-explanation effect: The 4 students who were subsequently more successful at solving problems at the end of the chapter (averaging 82% correct in the posttest) were the ones who *spontaneously* generated a greater number of self-explanations while studying the examples (15.3 explanations per example). The 4 less successful students averaged 46% correct on the posttest and generated only 2.8 explanations per example (Chi et al., 1989).

The self-explanation effect has been directly replicated in other laboratories, all in the domain of learning a procedural skill. Pirolli and Recker

(1994) used a similar design, with LISP coding as the task domain. Ferguson-Hessler and de Jong (1990) had subjects give protocols as they studied a manual on applications of principles of electricity and magnetism to the Aston mass spectrometer. Nathan, Mertz, and Ryan (1994) examined the effect of self-explaining on solving algebra word problems. All of these studies found the number of self-explanations correlated with problem-solving successes.

There are numerous other findings in the literature that can be interpreted to *indirectly* support the self-explanation effect. Most notable are the robust findings of Webb (1989). Citing 19 published studies on learning mathematics and computer science in small groups, Webb found that *giving* elaborate explanations was positively related to individual achievement, whereas *receiving* elaborate explanations had few significant positive relationships with achievement. Such findings are consistent with the self-explanation effect because the advantages gained by explaining to others and to oneself are comparable. In testing the effectiveness of using a computer-based analysis tool for instructional design in teacher training, Russell and Kelley (1991) serendipitously found that requiring the students to explain their design decisions dramatically and beneficially altered their understanding of course materials. In a completely different context, it has long been known that reading out loud to young children is beneficial for their intellectual, social, and emotional development. However, such intellectual stimulation is further enhanced if the parent engages the child in discussions of the stories (Dole, Valencia, Greer, & Wardrop, 1991). More remotely, the phenomena of the generation effect (Slamecka & Graf, 1978), the hypothesis generation effect (Klahr & Dunbar, 1988), and cognitive tuning set (Zajonc, 1960), may all be interpreted to be broadly consistent with the self-explanation effect.

Concurrently, there is also a general momentum in the science education literature toward talking, reflecting, and explaining as ways to learn, especially for challenging science domains. The whole idea of the new "talking science" approach (Hawkins & Pea, 1987; Lemke, 1990) is that students should learn to be able to *talk* science (to understand how the discourse of the field is organized, how viewpoints are presented, and what counts as arguments and support for these arguments), so that students can participate in scientific discussions, rather than just *hear* science. Alternatively, our interpretation of the benefit of talking science is the provision of self-explanations: a constructive inferencing activity. Although both of these views encourage the students to talk more, and both make the same prediction concerning the beneficial learning outcome, they do make diverse predictions with respect to other forms of constructive activity, such as diagram drawing. The talking science view focuses on learning the dis-

course of science, whereas the self-explanation view accepts the possibility that any form of constructive activity may be beneficial to some degree, even diagram drawing.

Here, we explore the generality of this self-explanation effect by extending it from skill acquisition to the learning of a coherent body of new knowledge. Moreover, this study shows that the beneficial effect of self-explanations can be achieved merely by prompting students to self-explain. Thus, this current study has five specific goals:

1. To see if the self-explanation effect can be generalized to a different (nonprocedural) domain, a different task, a different outcome measure, and a different age group.
2. To develop an assessment of understanding based on complex questions that were designed in a principled way, so that the questions can be differentiated on the basis of what kind of knowledge (directly encoded, integrated, or inferred) they tap, as opposed to the more standard method of simply grading the difficulty of questions.
3. To develop an analysis to capture what understanding is, in terms of the changes from the students' initial to their final mental models.
4. To address the processing characteristics of self-explaining that may make self-explanations especially amendable for theory revision.
5. To see whether the same improvement in learning can be achieved when students are merely prompted to self-explain.

The last goal is particularly important in light of the fact that the self-explanation effect could simply be interpreted as yet another indication of two distinct learning types, comparable to binary learning approaches suggested by many other investigators. Bereiter and Scardamalia (1989), for instance, identified two kinds of learners: high intentional and low intentional. Marton and Saljo (1976) likewise postulated two learning approaches: surface versus depth oriented. More recently, Ng and Bereiter (1991) suggested that learners differ in the type of goals they maintain: task-completion goals, instructional goals, and personal knowledge-building goals. It is also the case that spontaneous generation of self-explanations can be correlated with ability. Therefore, in order to claim that it is the activity of generating self-explanations that fosters greater understanding rather than some basic difference in learning style or ability per se, we must further show that this activity can promote learning when elicited, independent of ability.

This study therefore attempts to replicate the self-explanation effect with some important changes:

1. Using a different domain (a biological domain instead of a physics domain);
2. Using a different age group (eighth graders instead of college students);
3. Self-explaining expository text rather than worked-out solution examples;

4. Focusing on declarative understanding of concepts (as assessed by question answering and induction of function) rather than learning a procedural skill (as assessed by problem solving); and most importantly,
5. Prompting students for self-explanations rather than relying on spontaneous generation of self-explanations.

In addition, this study includes a control group of 10 students who were not prompted to self-explain; instead, they were given an opportunity to read the text twice, so as to roughly equate time on task with the prompted group. Moreover, prior knowledge and ability were assessed by a set of pretests and achievement measures.

To review, two groups of eighth-graders were asked to read a passage about the human ciculatory system, taken from a popular high school biology text. Prior to reading they were given a set of pretests to assess their prior knowledge, and after reading, they took posttests to assess their understanding. During the reading phase, the prompted students were asked to explain, after reading each sentence, what they understood about that sentence. The unprompted students were asked to read the passage twice so as to roughly equate time spent on the task. California Achievement Test (CAT) scores for both groups of students were used as a measure of their abilities.

METHOD

Materials

Analysis and Choice of a Biological Topic

Understanding the circulatory system requires what philosophers have referred to as "systematic" explanation (Haugeland, 1978), in that one must understand the "organized cooperative interactions" that occur within the system. Such cooperative interaction can be explained by the systematic interaction of the distinct components at all levels of the system. One way to specify this organized cooperative interaction is to decompose the circulatory system into its components and identify the structure, function, and behavior of each component. Each physical entity, such as the atrium, was operationally defined as a component. There are three kinds of "local" features to each component: the structure, function (or purpose), and the behavior. Take the atrium, for example. One structural (S) property of the atrium is that it is a muscular chamber; the function (F) of the atrium is to serve as a holding bin for the blood returning from the body or the lungs; the behavior (B) of the atrium is that it contracts and squeezes the blood into the ventricle. At this first level of analysis, textbooks can be inadequate

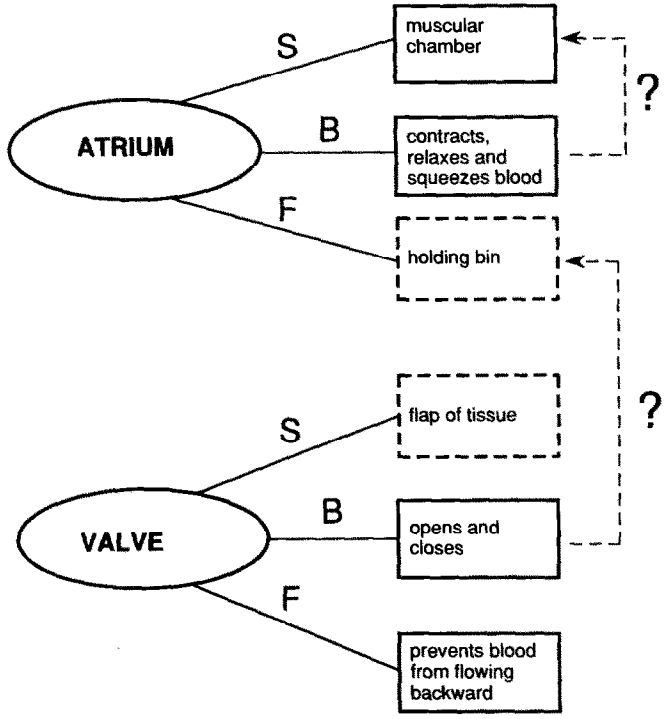

Figure 1. A schematic diagram of the knowledge pieces about components and their relations. Dotted lines and boxes indicate information missing from the text (S=structural property; F=function; B=behavior).

by omitting information about certain features of each component. Typically, the functional information is omitted. See Figure 1 for a depiction of the three local features of each component. Here, "local" is taken to mean features that pertain directly to that component.

There are relations within these three features of each component that could be specified in a textbook but are often omitted, such as the relationship between the structure of the atrium and its behavior (as indicated by the upper question mark in Figure 1). In addition to these within-a-component relational features that are left unspecified, there are also relations among components. For each feature of an individual component, one can ask about the relation between that feature (such as the function of the atrium as a holding reservoir) and a feature of another component (such as the behavior of the valves in the heart; see the lower question mark in Figure 1). If a learner understands this relationship, then the learner should understand the health consequence of having a murmur or a leaky valve. The textbooks at the junior and senior high school levels typically leave many of these within-a-component and among-components relations unspecified.

However, if a learner understands all the local features of each component, then, presumably, these relational consequences within and among components can be inferred. They are required for deep understanding.

So far, three kinds of information that could be specified in an expository passage have been mentioned: the local features of each component, the relationships among the local features of each component, and the relations among the local features of different components. Besides these, for the circulatory system at least, there are also hierarchical relationships between the local features of individual components and the systemwide features of the circulatory system as a whole. For instance, the systemwide goal of the circulatory system is to deliver nutrients and oxygen to the body's cells and to remove wastes. There are relations among the local features of a component with this top-level purpose. So, for instance, one structural feature of the heart is that it is divided in the middle by the septum. This feature has its own local purpose, which is to separate the oxygenated from deoxygenated blood within the heart. But this local function is subordinate to the systemwide purpose of optimizing the functioning of the circulatory system as a whole by allowing for separate pulmonary and systemic circulations. Thus, in order to have a coherent understanding of the circulatory system, an intricate system of relations must be understood not only locally, but systemwide as well. In addition to these kinds of within-a-component, between-components, and hierarchical relations, "processes" such as diffusion or oxygenation constitute a fourth kind of relationship involving several components.

All the information that is omitted from a text has been described here in great detail to illustrate three points. The first is that, although there have been general attempts to improve texts by inserting more elaborations and so on (Beck, McKeown, Sinatra, & Loxterman, 1991), such undertakings seem pretty hopeless because there are numerous relations that are left unspecified. Explicating all the relationships would make texts ridiculously long. The second point is that because information is omitted, any expository passage leaves a great deal of room for readers to provide their own inferences to bridge the gaps in the information provided. Hence, self-explaining seems to be a necessary activity in order to maximize what is learned from any expository passage. The third point is that the design of the questions for assessing understanding in this study was motivated largely by assessing these kinds of relational understanding.

Choice of Text Passage

An analysis of four highly rated texts corroborate the analysis as described in the preceding section.[1] It appears that, in general, most of the texts are

[1] We appreciate the help of Joanne Striley for this analysis.

more or less adequate in explaining the global function of the system, factual knowledge about the structure of components, and behavior of the components. Deficiencies (in terms of the incompleteness of the account) occur primarily in the other factors, such as the function of components, local relations within and between components, systemwide relations, and general processes such as diffusion. The four texts differed chiefly in the extent to which they relate anatomical structure or function to physiological functioning of the entire system. Because such an analysis shows that all texts have shortcomings of one kind or another, the final choice was to use Towle's (1989) *Modern Biology*, a highly rated and popular text, usually used in the ninth grade.

Towle's (1989) unit on the circulatory system was kept intact, except that the three figures (not particularly good ones) and a few sentences were deleted to keep the passage short. The edited version contained a total of 101 sentences, relating to the major components of the primary topic. Sentences that were deleted either referred to figures in the original textbook or were excursions into related topics (such as the lymphatic system), which were not mandatory for understanding the primary topic. Figures in the text were deleted so as not to complicate the interpretation of the results with differences in the ability to learn from diagrams (Mayer, 1993).

Consistent with other good texts, this one also contained information primarily about the global function of the circulatory system, as well as the structure and behavior of each component. This unit explicitly discussed 11 of the 22 components' functions. Even when the local function of a component was explicitly specified, it was usually not related to the systemwide function of the circulatory system. For instance, the passage would indicate the purpose of the pulmonary system and the structure of valves in the veins. However, it would not state why the pulmonary vein does not have a valve in it. Therefore, the students would have to make an inference (or a chain of inferences) in order to complete their understanding.

How the Questions Were Designed

The design of the set of questions used in the pre- and posttests was intended to focus on tapping new knowledge that the reader had to construct from the information presented. This constructed new knowledge often corresponded to the links that are depicted by the question marks in Figure 1. To devise a way of designing some of the questions, each sentence of the passage was coded as to what type of information it contained, whether it described the function, and/or the structure, and/or the behavior of a component, and/or relationships between components. Occasionally, a sentence was coded as a "factlet," which was a sentence that contained none of the three local features or any relations among the components, such as "The heart continues to beat without interruption more than 2.5 billion times in the average life span" (Towle, 1989, p. 655). Processes such as dif-

fusion generally required several sentences to describe. The purpose of such coding was to design questions in a principled way, so that one could predict exactly what knowledge the question was assessing, and whether this knowledge could be directly gained from a sentence, or whether it was inferred. For instance, if a sentence described only the structure of a component, such as the atrium, could a learner answer a question about its function?

Four sets of questions were constructed: Three sets of questions (Categories 1, 2, and 3 in ascending difficulty) were designed to test what was learned from the passage, and Category 4 (or health) questions tested students' use of this newly acquired knowledge to answer health-related questions.[2] How these questions were constructed now follows.

Category 1 contained 14 "verbatim" questions that were generated from information explicitly stated in the text. Usually, the information was presented in a single sentence but, occasionally, an implied agent from the previous sentence was needed. Thus, these questions asked about the structure, function, or simple knowledge about processes, as directly or explicitly presented in the text. Table 1 contains an example of a Category 1 question along with the sentence from the text on which that question was based.

Category 2 contained 14 "comprehension inference" questions based on material explicitly presented in the text. However, these questions required the student to integrate information from two or more lines of text, or to integrate across nonconsecutive paragraphs. For example, one sentence in the passage described the structure of the skeletal muscles, and another sentence described the behavior of blood flow. A question was then posed concerning the relationship between the structure of the skeletal muscles and the direction of blood flow (see Table 1 for an example). Notice that most of the work in the comprehension literature assesses understanding by this sort of comprehension-inferencing questions (e.g., see the questions designed by Goldman & Duran, 1988; Graesser & Murachver, 1985; Nicholas & Trabasso, 1980; Nix, 1985), in that the questions require either paraphrases, explicit comparisons, or integration of information within and across paragraphs. Thus, these questions require comprehension inferencing of varying degrees, but all the knowledge needed to make the inferences is presented in the sentences.

In contrast, Category 3 contained 14 "knowledge inference" questions that required the *generation* of new knowledge. Answering questions required a good deal of understanding, and the use of prior knowledge (either domain-relevant, commonsense, or everyday knowledge). These questions were of many types. The first type assessed whether a learner could induce

[2] Two additional categories of questions, pertaining to medieval misconceptions, and to scientific discoveries, were also designed and administered for different purposes. These questions were intended to compare and contrast contemporary and historical misconceptions, and to examine the inferential processes in discovery. Because the questions were not designed to test the self-explanation effect, no analyses of them will be reported.

TABLE 1
Examples of Category 1–4 Questions
and the Corresponding Lines of Text Needed to Answer Them (Towle, 1989)

Category 1: Verbatim	
Sentence 127:	Hemoglobin is the molecule that actually transports oxygen and carbon dioxide.
Question:	*What does hemoglobin transport?*
Category 2: Comprehension Inferences	
Sentence 71:	Many veins pass through the skeletal muscles.
Sentence 72:	During movement, these muscles contract, squeezing blood through the veins.
Question:	*Besides the role of valves in the veins, what keeps blood from the lower parts of the body moving up toward the heart through the veins, against gravity?*
Category 3a: Knowledge Inferences	
Sentence 17:	The septum divides the heart lengthwise into two sides.
Sentence 18:	The right side pumps blood to the lungs, and the left side pumps blood to other parts of the body.
Sentence 30:	Blood returning to the heart, which has a high concentration of carbon dioxide and a low concentration of oxygen, enters the right atrium.
Sentence 33:	In the lungs, carbon dioxide leaves the circulating blood and oxygen enters it.
Sentence 34:	The oxygenated blood returns to the left atrium of the heart.
Question:	*Why would the distribution of oxygen* (a systemwide function) *be less efficient if there is a hole in the septum* (a structure of the septum)*?*
Category 3b: Knowledge Inferences	
Sentence 73:	Valves prevent the blood from moving backward or downward.
Sentence 80:	Pulmonary circulation is the movement of blood from the heart to the lungs and then back to the heart.
Sentence 86:	Oxygenated blood then flows into venules, which merge into the pulmonary veins that lead to the left atrium of the heart.
Sentence 87:	The pulmonary veins are the only veins that carry oxygenated blood.
Question:	*Why doesn't the pulmonary vein have a valve in it?*
Category 4: Health	
Sentence 3:	The blood moving through the vessels serves as the transport medium for oxygen, nutrients, and other substances.
Sentence 64:	Several venules in turn unite to form a vein, a large blood vessel that carries blood to the heart.
Sentence 76:	The English scientist William Harvey first showed the heart and blood vessels formed one continuous, closed system of circulation.
Question:	*Some snake bites can be dangerous because the snake venom causes muscle paralysis (muscles become immobile—can't move). How is it that a person can die in a short amount of time from such a snake bite, even when the bite is on the ankle?*

the function of a component if only the structure or behavior was explicitly described in the text. This went beyond just asking students in a straightforward way what the function of a component was: They had to induce the function in order to use that knowledge to answer a question that was not directly related to the function.

A second type of question in this category required students to induce the function (or behavior or structure) of a component and relate it to another feature (function, structure, or behavior) of another component. For example, suppose the behavior of the valve was described as opening and closing. A question then asked about the relationship between the opening and closing of valves and the purpose of the atrium.

A third type required relating a specific induced feature of a component to either a systemwide purpose or some process such as diffusion. A question of this type is shown in Table 1, Category 3a.

> *Why would the distribution of oxygen* (a systemwide function) *be less efficient if there is a hole in the septum* (a structure of the septum)*?*

In order to answer such a question, the student needs to induce the function of the septum because none of the relevant sentences in the text (Sentences 17, 18, 30, 33, and 34) explicitly mention it.

A fourth type of question asked for an explanation of an implied structure. Table 1, Category 3b, illustrates this type of question. All the sentences (73, 80, 86, and 87) from the passage that were relevant to valves and the pulmonary vein are shown in this table. No sentence mentioned anything about the role of valves in the pulmonary vein. Hence, a question relating these two components asked why the pulmonary vein would not have a valve in it. In order to answer such a question, the student must reason about and infer several notions: First that most veins have low pressure, but that the pulmonary circulation is only a short distance away from the heart, therefore the heart can pump blood to the lungs with a great deal of force, thereby eliminating the likelihood that the blood will flow backward, hence no valve is needed. A great deal of the reasoning brings in commonsense knowledge, such as the shorter distance of the pulmonary circulation implies greater pressure. This factor is one of the reasons the conjecture is made that self-explaining can facilitate understanding, because it encourages the student to utilize commonsense knowledge.

Finally, Category 4 contained 14 "health" questions that typically assessed students' understanding of the implications of the systemwide properties of the circulatory system. One such systemwide property is that the circulatory system is a closed system of circulation. A closed system implies certain properties such as, the same blood circulates throughout the system, and the total volume of blood does not change. The snake bite question (shown in Table 1) assesses students' use of such knowledge.

The pre- and posttests also included two other tasks: drawing the pattern of blood flow on an outline of a human body, and defining 23 terms that were taken from the unit.

Subjects

Eleven pilot students were tested prior to the actual study. These pilot students were tested for a variety of purposes, such as modifying the loci of probes in the text for knowledge of functions, adding definitions of a couple of vocabulary words used in the passage that students did not understand, and timing the duration of reading the entire passage with prompting. The experimental subjects were 14 eighth graders (7 boys and 7 girls) for the prompted group and 10 eighth graders (5 boys and 5 girls) for the unprompted group, all recruited from a local public school. More students were tested for the prompted group with the intention to do some contrastive analyses. It would have been preferable to test more students in total, but the subject sample was restricted to volunteers from all the eighth graders of an inner city school class (a small sample to begin with) in order to maintain homogeneity. None of them had taken a biology course.

Students were chosen intentionally with a range of abilities in terms of their CAT scores, so that ability differences could be examined. Even though students were paid for their participation, it was difficult to recruit students with lower CAT scores because they tended not to volunteer for this kind of activity.

Procedure

There were three phases to the study. First, an initial interview session (a pretest) in which each student discussed what they knew about 23 terms of the circulatory system, then drew the direction of blood flow through the body and answered half of the questions in each category (1–4), in that order. This interview was audiotaped. More specifically, in the first section of both the pre- and posttests, students were asked to explain everything they knew about 23 terms taken from the circulatory system (i.e., heart, lungs, pulmonary artery, etc.). They were asked to consider:

1. What is it? What kind of thing is it? What does it refer to?
2. Where is it found in the body?
3. What is its structure, texture, or composition?
4. What does it do?
5. What is its purpose?

In the blood path section, students were presented with an outline of the human body and asked to draw the path of blood flow throughout the

body, making sure the path included the heart, lungs, brain, feet, and hands.

Second, each student then read the 101-sentence passage. Each sentence was printed on a separate piece of paper. The prompting consisted of general instructions (shown in the Appendix) given at the beginning of the reading phase. Then the students were told to explain what each sentence means. The students had absolutely no trouble carrying out this general instruction. The prompt given by the experimenter after every sentence was very general, more analogous to a reminder. In fact, after a few initial prompts, the students often proceeded with self-explaining without any further prodding, perhaps because the one-sentence-per-page format served as a sufficient cue.

In addition to these general prompts, a set of specific function prompts were inserted at 22 locations throughout the 101-sentence passage. These locations corresponded to places in the text at which a component of the circulatory system was discussed, such as the atrium. The function prompt occurred after the student generated the self-explanation of that sentence, and consisted of asking the student explicitly what the function of the component was.

Besides the explanations they generated in response to reading each sentence, students may have been prompted for further clarification by the experimenter if what they stated was vague, with comments such as "Can you explain that?", "What do you mean?" or "Flowing through what?" (if the student said "Flowing through there"). Students were told that they could also take notes and draw diagrams while reading, although they were not prompted to do so in any systematic fashion.

Students in the control or unprompted group were not prompted for either self-explanations or functions. Instead, they were asked to read the materials twice. Reading it twice (as opposed to three or four times) engaged the unprompted students with the text for roughly the same amount of time as having the prompted students read and explain the text. This was determined a priori by measuring the total amount of time a pilot group of prompted students took. The actual mean of the prompted group's studying time was 2 hr 5 min (range = 1 hr 27 min to 2 hr 53 min), and the mean of the unprompted group was 1 hr 6 min (range = 22 min to 2 hr 47 min).

Third, after they read the material, students took a posttest, which included the 23 terms, the blood path, and the entire set of the Category 1–4 questions, including those given in the pretest. In order to reduce stress and place more emphasis on learning as opposed to memory, students were allowed to look back at the text or their notes to help explain any terms or answer any questions in the posttest.

The whole precedure was run in three to four sessions (some students took two sessions to read the text), with each session spaced at least a week apart. Each session lasted from 1–3 hr. All sessions were audiotaped.

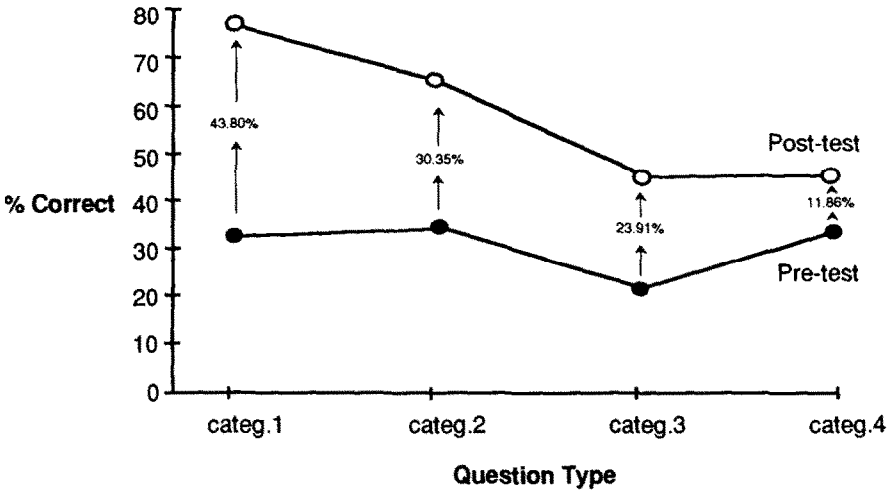

Figure 2. Percentage of correct answers on pre- and posttests across all 24 subjects for all four categories of questions (all gains are significant at the .001 level).

RESULTS

Overall Gains

Student's answers on all four categories of questions were scored by matching answers to a template of idealized answers constructed by the researchers. Each question had a maximum score of 6 points, with partial credit available depending on the question. Collapsing across the two groups, Figure 2 shows significant gains in all four categories of questions ($p < .001$), suggesting that both groups of students took this learning task seriously, and were able to retain the knowledge for at least a week.

Differences Between the Prompted and Unprompted Groups

Both prompted and unprompted students gained significantly greater understanding from the pretest to the posttest, $F(1, 22) = 183.6$, $p < .0001$ (see Figure 2), and this gain was more pronounced for some categories of questions than others, $F(6, 132) = 24.7$, $p < .0001$ (see Figure 2 again). However, from the pretest to the posttest, the gain was greater for the prompted group (32%) than the unprompted group (22%; see Figure 3). This interaction, between groups and tests, is significant, as shown by a $2 \times 2 \times 4$ (Prompted vs. Unprompted $\times$ Pretest vs. Posttest $\times$ Question Categories) analysis of variance (ANOVA) with repeated measures, $F(1, 22) = 5.1$, $p < .05$. There was no significant main effect for the groups, $F(1, 22) < 1$, as there were for tests and question categories, as noted before. The three-way interaction was not tested because categories were nested within tests.

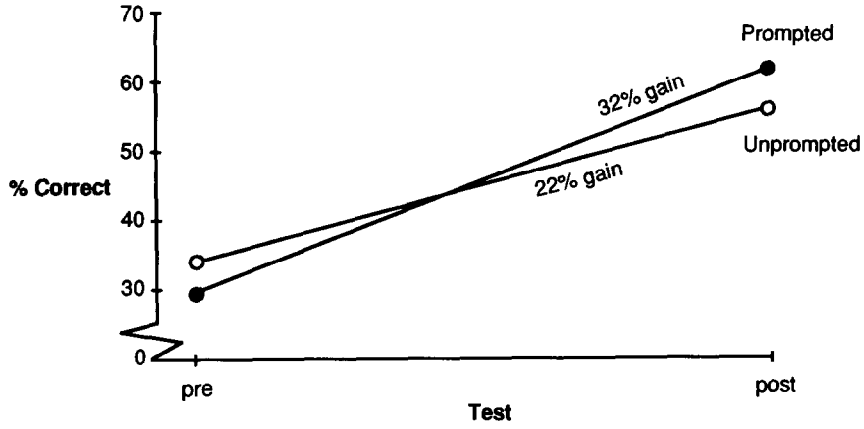

Figure 3. Percentage correct answers on pre- and posttests comparing prompted ($n=14$) and unprompted ($n=10$) students.

The difference between the two groups' improvement is even more dramatic if only the more difficult Category 3 and 4 questions are examined: The prompted group improved by 22.6% versus only 12.5% for the unprompted group, $t(22)=2.64$, $p<.01$. These more difficult questions required knowledge inferences and use of commonsense knowledge, and were apparently more sensitive measures of the improvement afforded by self-explanations.

The greater gain of the prompted group, especially for the more complex questions, is particularly impressive considering the following. First, the selected passage from Towle (1989) is already a very well written text; even so, self-explaining can further enhance comprehension. Second, the students from the unprompted group were given the opportunity to read the passage twice, thus reinforcing their understanding of the text. Nevertheless, the prompted group still outperformed them. Third, the unprompted group was *not* suppressed from self-explaining, so that some of the students may have spontaneously self-explained anyway while reading the materials twice (some evidence of this will be provided later). Recall that the original self-explanation effect (Chi et al., 1989) was obtained by contrasting spontaneous with nonspontaneous self-explainers. In other words, the contrast is *not* between explainers and nonexplainers as this analysis does not capture the extent to which the prompted and the unprompted students were self-explaining. The data merely show the additional improvements gained by prompting students to self-explain. Finally, the greater gain was obtained by merely eliciting self-explanations; no extensive training program was undertaken.

The prompted group did have higher CAT scores than the unprompted group ($M=87\%$ vs. $M=83\%$), but the difference was not significant. Using

the gain scores (i.e., the difference between the pre- and posttests scores) as a dependent measure, the influence of ability on prompting outcomes was examined with multiple regression. CAT scores were not a significant predictor in the regression analysis, $t(21) = 1.479$, $p = .15$. With CAT scores included, the prompted group retained a significant advantage, $t(21) = 2.147$, $p < .05$. Whereas these results do not rule out the role of prior ability in the outcomes, it is clear that ability had a much smaller influence on a student's improvement than whether they were assigned to the prompted group. Additional analyses on ability, presented later, are consistent with this interpretation.

Within the Prompted Group Analyses

Codings of Self-explanations

Although it is clear from the preceding analyses that prompting is successful at eliciting self-explanations and enhancing learning, the effect of the prompting intervention could have been more pronounced had it been possible to ascertain the extent to which the unprompted group did or did not self-explain covertly. Because the previous comparison between the prompted and the unprompted group was not a contrast between explainers and non-explainers, it may be more productive to know exaclty how many self-explanations each student in the prompted group generated and contrast the high explainers with the low explainers. (Note that no analysis can be carried out with the unprompted group because they did not overtly self-explain.)

In order to determine the amount of self-explanations each student in the prompted group generated, it was necessary to define and code what a self-explanation was. Self-explanation in the physics work (e.g., Chi et al., 1989) was operationalized as any utterance that went beyond the information given, namely, an inference of new knowledge. Suppose Student 4 read the following sentences about the septum:

> Sentence 17: The septum divides the heart lengthwise into two sides.
> Sentence 18: The right side pumps blood to the lungs, and the left side pumps blood to the other parts of the body.

Ignoring what the student said after Sentence 17, the explanation generated after Sentence 18 was:

> So the septum is a divider *so that the blood doesn't get mixed up*. So the right side is to the lungs, and the left side is to the body. So the *septum is like a wall* that divides the heart into two parts...it kind of like *separates* it so *that the blood doesn't get mixed up*....

There are two pieces of knowledge inference here: The fact that blood in the two sides does not mix, and that the septum is solid like a wall. These pieces of knowledge are crucial for understanding, and many historical and every-

TABLE 2
Example of Self-explanation Coding

Sentence 118:	These substances (including vitamins, minerals, amino acids and glucose), are absorbed from the digestive system and transported to the cells.	
Student:	"Okay, so that's the point of what hepatic portal circulation [is]. To pick up these nutrients, I would guess."	Explanation 1
Experimenter:	"Okay, why would you say that?"	
Student:	"Well because it says that it's absorbed from the digestive system, um vitamins, minerals, amino acids, and glucose,	Paraphrase
	and so that's why it's important to eat a balanced diet, or else your cells won't get the right vitamins, minerals amino acids, and glucose."	Explanation 2
Explanation 1:	**Purpose of the hepatic portal circulation: to pick up nutrients from digestive system.**	
Explanation 2:	**Eating a balanced diet is important for your cells.**	

day false beliefs pertain to these ideas, such as that the septum is porous and permits the blood from the two sides to exchange. These are inferences that make use of commonsense knowledge, such as that dividing lengthwise can mean separating like a wall, which also implies that it may be solid.

Self-explanations were considered to be inferences that went beyond the text, *excluding* monitoring statements, paraphrases, comprehension, or bridging inferences (these were considered kinds of paraphrases). Comprehension inferences were either whole-sentence paraphrases, or direct translation of individual words (e.g., such as translating the word "transport" to the word "carry," or translating the word "divides" to the word "separates," as shown in the previous self-explanation quote), or explicating implicit references, and so forth. Table 2 shows an example of two utterances coded as self-explanations in contrast to one coded as a paraphrase. (The underlined text is what the student read. The explanations uttered by the student are in quotation marks. The recoded content of each explanation follows.)

The identified comments in Table 2 were counted as two units of self-explanations, as indicated. Hence, basically, parts of the protocol utterances were coded as knowledge inferences if the coder, who was a project member, subjectively felt that new pieces of knowledge had been generated,

TABLE 3
Example of Coarse and Fine Grain Coding of Self-explanations

Sentence 9:	During strenuous exercise, tissues need more oxygen.
Student:	"During exercise, the tissues, um, are used more, // and since they are used more, they need more oxygen and nutrients. // And um the blood, blood's transporting it [O_2 and nutrients] to them. //"
Coarser Coding	
Explanation 1.	Purpose of blood is to transport oxygen and nutrients to the tissues.
Finer Coding	
Explanation 2.	Tissues are used more during exercise.
Explanation 3.	Since tissues are used more, they need more oxygen.
Explanation 4.	Purpose of blood is to transport oxygen and nutrients to the tissues.

as illustrated before with respect to the "blood doesn't get mixed up" and "septum is like a wall" examples. Notice, also, that in that quote, even though the student mentioned the fact that "blood doesn't get mixed up twice," only one explanation inference was counted. This coding produced a fairly coarse grain size, which is similar to the grain size used in the original Chi et al. (1989) data.

This coarse level of coding might overlook more minute inferences that a student could have generated in those uncoded parts of the protocols. Hence, for a second pass, another project member systematically coded *all* the utterances generated in response to reading each sentence, into units that corresponded more or less to a proposition. Not surprisingly, an explanation could span several propositional units. The segmentation into propositional sized units forced us to systematically consider whether every individual proposition was an inference, and if not, whether it participated in the context of a larger sized inference. This subsequent coding was extremely tedious, and took an additional 200 person hours. Naturally, a finer level of coding tended to capture more inferences at a smaller level. Table 3 illustrates and contrasts the original coarser coding with the second finer coding. It shows the sentence the student read (the underlined sentence), followed by her explanation of the sentence (in quotation marks). The slash marks (//) indicate where proposition-like segments were coded.

Although the finer coding captured more inferences (3 vs. 1, see Table 3), the overall pattern of results for the two sets of codings was largely identical. Therefore, the two sets of codings serve as a validity check. The majority of the analyses in this article were then carried out with the coarser coding for three reasons. First, pragmatically, the coarser coding reduced the amount of work for many of the subsequent analyses. Second, the self-explanations captured at the coarser level made more intuitive sense as a piece of knowledge inference. Third, the finer coding sometimes seemed to capture comprehension inferences (as in the case of Explanation 2 in Table 3) and

redundant inferences (such as Explanation 3 in Table 3). On the other hand, because the finer coding considered all the utterances, it was necessary as a measure of the amount of talk, so that the effect of verbosity could be assessed.

Outcome Measures

High and Low Self-explainers. The previous contrast between the prompted and the unprompted groups did not give us a sense of how many explanations each group did generate, especially because the unprompted students could have spontaneously generated them covertly. Coding the prompted students' explanations into units tells us precisely the number of self-explanations each student did generate, so that we can contrast the performance of the high explainers with the low explainers.[3] Therefore, the students were ranked according to the number of self-explanation inferences they generated (see Table 5 under the "Inferences" column). To maximize the contrast, the top 4 high explainers (Subjects 1–4) were compared with the bottom 4 low explainers (Subjects 9–12), not counting the 2 lowest subjects (because Subject 14 had an outlying CAT score, and Subject 13 had no available CAT score). The high explainers generated a mean of 87 self-explanations and the low explainers generated a mean of 29 self-explanations across the entire passage using the coarser coding, $t(6) = 5.64$, $p < .001$. The difference remains significant even if the incorrect self-explanations are removed. That is, the high explainers ($M = 61$) articulated a greater number of correct explanations than the low explainers ($M = 14$), $t(6) = 5.89$, $p < .001$. A similar difference was obtained by using the finer grained coding. High explainers generated a mean of 117 self-explanation propositions versus a mean of 60 for the low explainers, $t(6) = 4.42$, $p < .01$.

The high explainers' pre- to posttest gain scores improved more than the low explainers (38% vs. 27%), although this difference approached significance, $t(6) = 2.14$, $p < .07$. (It would be significant if the correct answers derived from subsequent looking up in the text passage were excluded, as will be shown later.) Again, as was the case contrasting the prompted and unprompted groups, the improvement was more pronounced for Category 3 and 4 questions (33% vs. 17%), $t(6) = 3.02$, $p < .05$, which suggests that they acquired a deeper understanding of the topic (see Figure 4).

Recall how knowledge inference and health questions were constructed. These were the most complicated questions, tapping knowledge of system-wide features (such as closed and double circulation), and asking for a

[3] In the interest of seeing whether the current results replicate the previous finding in the domain of physics (Chi et al., 1989), instead of contrasting high and low explainers, one can also contrast the more and less successful learners, splitting the subjects on the basis of their question-answering scores, comparable to the successful and less successful problem solvers in physics. An identical pattern of results was obtained.

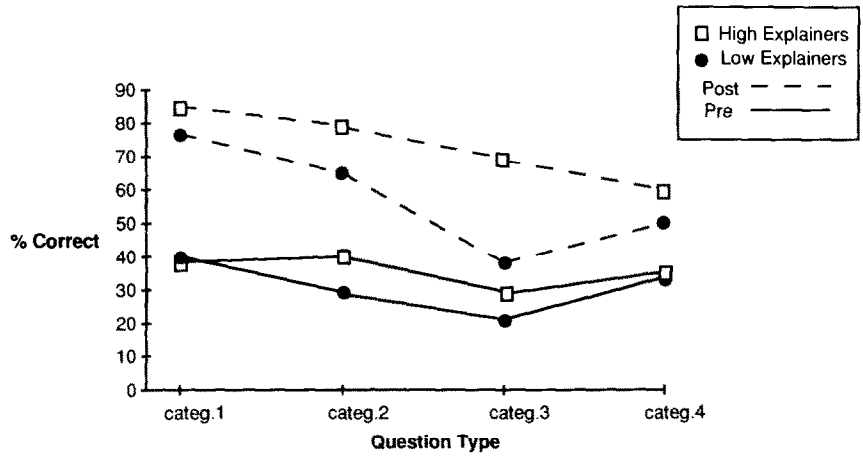

Figure 4. Percentage of correct answers for Category 1–4 questions on pre- and posttests comparing high explainers ($n=4$) and low explainers ($n=4$).

causal explanation of an implied structure. Another way to think about the nature of these questions is that the majority of them required an inference of either the function, the structure, the behavior of a component, and/or the relation between the inferred structure, function, and behavior with another component's structure, function, and behavior. The structure of these knowledge inference and health questions can be seen by coding the questions in the same way that each sentence of the passage was coded. Such an analysis emphasizes the number of inferences that a student needs to construct in order to be able to answer these questions. Because high explainers answered significantly more Category 3 and 4 questions correctly than low explainers, it is evident that they *induced* what was only implicitly presented (as in the case of the question shown in Category 3a of Table 1, where the student must induce the function of the septum), and *used* the inferred knowledge to answer questions appropriately.

Because the question-answering task was an open-book one, students were free to refer to the text passage if they wished. High explainers, on average, referred to the text for only 2 of the 54 posttest questions, whereas low explainers, on average, referred to it for 11 of the questions. This suggests that the high explainers absorbed more understanding from reading the text with self-explanations, and did not need to reference it again. If the correct answers due to looking up information in the text are excluded from the analysis (mostly in the case of low explainers), the difference in pre- to posttest gain scores between high and low explainers becomes even more pronounced (37% gain for high and 15% gain for the low explainers when looked up answers are excluded), $t(6)=3.22$, $p<.02$.

Interestingly, only one of the high explainers' references regarded answers for Category 3 and 4 questions, whereas about 35% of the low explainers' text references were for Category 3 and 4 questions. This suggests that the high explainers may have known that knowledge inference and health questions did not contain answers that one could look up in the text. These findings on text referencing are consistent with the example-referencing behavior of the more and less successful problem solvers in the Chi et al. (1989) study, in that the more successful solvers made fewer references to examples and reread fewer lines per example, suggesting that they knew exactly what they needed to look up.

Using the finer grained analysis, the high explainers did speak more than the low explainers (650 propositional units, $SD = 198$, vs. 439 propositional units, $SD = 71$, respectively), but the difference was not significant. Many factors contribute to the source of variance for verbosity: the number of repetitions, irrelevant comments, metastatements, paraphrases, and so forth. For instance, some students repeat themselves more often than others, as seen earlier in the quote of Student 4's explanation. Although it is difficult to partial out the effects of general verbosity from generating self-explanations, it is true that the high explainers had a higher proportion of self-explanation propositions ($117/650 = 18\%$) than the low explainers ($69/439 = 14\%$). This is at least suggestive that the high explainers were doing more than just talking more.[4]

One characteristic of self-explaining is that it is a constructive activity. Naturally, one wonders whether other forms of constructive activity could also enhance learning. This notion can be tested by comparing the effectiveness of generating self-explanations with drawing diagrams in the process of self-explaining. Recall that the students were encouraged (but were not

[4] One question that often arises is the legitimacy of considering self-explanations as a source of verbal protocols because the process of generating self-explanations alters the processing of the to-be-learned materials. This is to be contrasted with protocols collected while a subject solves a problem, as reported in Ericsson and Simon (1984), which presumably does not alter the processing of the task. The difference is that in the traditional protocol collection context, subjects are only supposed to articulate the information passing through their short-term or active memory as they are solving a problem. Verbal protocols in that context are used to capture the processing sequences (of problem space states and operators applied) that the solver is undertaking. The solvers are prohibited or discouraged from any kind of analyses or reflection of the content of their short-term memory. Indeed, reflection on problem solving (retrospective protocols) is not a reliable indicator of processing (Ericsson & Simon, 1984), in part, because self-explanations may occur. Only concurrent protocols are considered reliable traces of problem-solving processes. Self-explaining (i.e., generating new knowledge), whether prompted or spontaneous, in contrast, is a process of reflection. Students are encouraged to reflect, think about, and infer what they are reading. In sum, the process of giving protocols is just diplaying the problem solving, whereas the process of generating explanations is adding inferences and constructing a mental structure. For a more extended discussion and manual on how to carry out alternative methods of verbal analyses, see Chi (1994).

systematically prompted) to draw diagrams and take notes if they wished. Diagrams included drawings of the circulatory system, often showing the interrelationships among the components, as well as treelike diagrams of concepts, but excluded notes, which were often copying of the text sentences. High explainers drew, on average, 18 diagrams ($SD = 22.0$) versus 1 diagram ($SD = 1.4$) for the low explainers, but the difference was not statistically significant. There was a great deal of variability in the number of diagrams drawn among the high explainers even though there was very little variability among the low explainers. All of the students who drew many diagrams were high explainers. However, the converse is not true: Not all of the students who drew few diagrams were low explainers. This suggests that drawing diagrams may be an alternative constructive activity for enhancing learning, so that the benefit of talking science may be its constructive nature, rather than the learning of the discourse and argument structure of science. However, it is not clear whether drawing diagrams alone would have been adequate at promoting greater learning, or whether it merely accompanied the activity of self-explaining. One of the problems is that the number of data points is slim compared to the number of self-explanations, so that drawing diagrams is not as sensitive a measure as self-explanations. It needs to be studied more systematically.

As an aside, because all frequent diagram drawers were high explainers, we can use the frequency of diagram drawing to diagnose whether students in the unprompted control group were covertly self-explaining. It turns out that 2 of the 10 unprompted students drew 21 and 17 diagrams, suggesting that *at least* 2 of the unprompted students were covertly spontaneously self-explaining.

High- and Low-Ability Students. One important question to ask regarding ability is the issue of aptitude–treatment interaction, that is, can lower ability students profit just as much from self-explanation prompting as higher ability students? By selecting 5 explainers with the highest CAT scores (98%) and 5 explainers with the lowest CAT scores (72%), two groups of explainers with significantly different ability differences were attained, $t(8) = 3.33$, $p < .01$. Their pre- to posttest gain scores on Category 1–4 questions shows that both higher and lower ability students profited the same amount, with gains of about 30% (see Figure 5).

The Content and Structure of Self-explanations. One can analyze the content of self-explanations in order to ascertain how they are produced. In the Chi et al. (1989) physics work, the content of self-explanations was captured by translating all of the self-explanations into *if–then* conditional statements, and then classifying them into one of four categories: systems, technical procedures, principles, and concepts. From this coding, each individual *if–then* conditional statement was determined to have been deduced

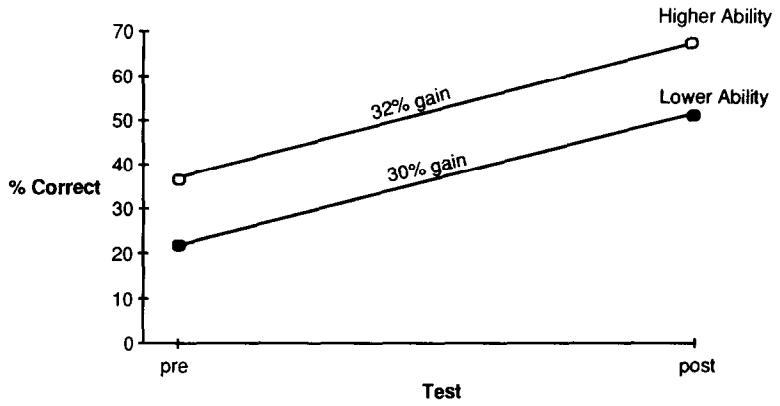

Figure 5. Percentage of correct answers on pre- and posttests comparing higher ability ($n=5$) and lower ability ($n=5$) prompted students.

from information presented earlier in the prose parts of the text or from information contained in the examples themselves. In general, those results (Chi & VanLehn, 1991) showed that self-explanations were minute inferences that were inferred either from the example lines, or else from integrating what was encoded from the example lines with commonsense and other background knowledge.

The picture that has emerged from the analyses of the structure of self-explanations in biology is essentially the same. It was determined how each self-explanation inference could have been generated, that is, with what source of knowledge was the inference generated, such as by integrating the currently read line of text with previously read line(s) of text, or else by integrating the currently read line of text with some commonsense or background knowledge. Explanation 2 in Table 2 (eating a balanced diet) is an example of a self-explanation that integrates newly presented information with commonsense knowledge. Likewise, analogizing the septum to a wall (quoted earlier) is also an example of the use of everyday knowledge. It was found that 30% of the self-explanations were produced by integrating new information with commonsense and background knowledge; 41% of the self-explanations were integration of new information with prior sentences; and the other 29% fell into various types, such as integrating with episodic experiences, using analogy, translating words or phrases, logical inferences, indeterminate ones, and so forth. In general, no differences in the distribution of the structure of these three kinds of self-explanations were found. That is, both high and low explainers generated self-explanations by integrating with prior knowledge at least 30% of the time because many of the ambiguous ones (29% of them) could be liberally classified as uses of prior knowledge.

Because self-explanation inferences are produced in around 41% of the cases by integrating the currently read line of text with previously read sentences, one can distinguish between links that integrate either across or within the same topic area (such as the structure of the heart, circulation in the heart, blood vessels, etc.). Hence, the number of references made to a sentence within the same topic area versus the number between topic areas was tabulated. For the high explainers, there was a 2:1 ratio in the number of integration within topics ($M = 34$ links) and between topics ($M = 17$ links). Low explainers, on the other hand, had a 5:1 ratio in the number of integration within topics ($M = 18$ links) versus across topics ($M = 3$ links). The fact that half of the high explainers' integrations were across topic areas suggests that they are more likely to produce a more coherent mental model in which distinct components of the mental model would be related to each other. Hence, the differential pattern of integration between the high and low explainers not only suggests a fruitful way to prompt students to generate utterances that constitute an explanation inference, but also may have implications for the coherence of their mental models, as will be shown later.

More Direct Assessment of Understanding

Induction of Function

Of all the outcome measures discussed in the preceding section, only one of them assessed the depth of understanding, namely, the high explainers' greater advantage in answering the more complex Category 3 and 4 questions than the low explainers because of the way these questions were constructed. Another explicit measure of understanding is provided by a test of how well students learned the function of components. Students' difficulty in grasping domains such as the human circulatory system is often blamed on texts that do not adequately supply information about the function of each component. In our passage, there were 22 identifiable components, of which only 11 had explicitly stated functions, so that the functions of the 11 others were implied. Thus, as a measure of understanding, all the students in the prompted group were explicitly probed for their explanations of the function of each of the components either right after or shortly after the text's introduction of that component. For example, the septum was introduced in Line 17 of the text. Student 4 was probed for the function of the septum after she had a chance to read and explain Sentence 18, as shown earlier see sentences 17 and 18 on Table 1. After Student 4 have her explanation, the experimenter probed her with the question:

Experimenter: *Can you explain to me what you think the purpose of the septum is?*

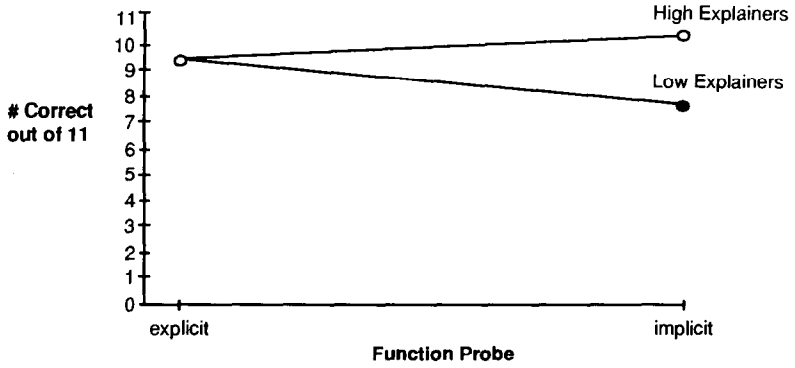

Figure 6. Number of correct answers to explicit and implicit function probes comparing high ($n=4$) and low ($n=4$) explainers (the interaction is significant at the .01 level).

Student: Well, I basically think that the purpose is to separate the right chamber, the right part of the heart [from] the left part of the heart, because each one has a separate function, and so the right side pumps to the lungs, the left is to the body. So you need something to separate them so [that blood] won't mix, and that's the septum's job.

On the whole, students could correctly articulate 84.6% of the explicitly presented functions, and 73.2% of the implicitly presented functions. The extremely high performance in inducing the function answers a theoretically intriguing question about whether it is possible to induce the function of a system or component when only the behavioral and structural information is given (see, e.g., the work of deKleer & Brown, 1983). These data suggest that it is quite possible for students to induce the function when it is not specified.

Moreover, when the function was *explicitly* described, both the high and the low explainers could articulate it equally well when probed (see Figure 6). However, when the function was only *implicitly* stated (or not stated at all), the high explainers were able to *induce* it significantly better than the low explainers (10.5 vs. 7.8 of 11 probed), $t(6) = 3.05, p < .05$. This result accounts, in part, for the superior performance of the high explainers in answering knowledge inference questions because many of those questions required the induction of a function that was not explicitly presented, as illustrated earlier with the question about the septum. Thus, not only do we have evidence that prompted students actually *can* induce the function when it is not explicitly stated (thereby suggesting that one way to minimize the inadequacy of texts is to practice self-explaining rather than revising the texts),

but the high explainers' greater facility at this induction, along with their increased advantage in answering knowledge inference and health questions, further supports the interpretation that high explainers gained greater understanding than the low explainers.

Analyses of Mental Models

Instead of a piecemeal and cross-sectional look at the display of understanding, as afforded by averaged performance measures across students such as the mean number of Category 3 and 4 questions answered correctly, or the mean number of implicit functions induced, an integrated assessment of the changes in the representation as a function of understanding was needed. With this goal in mind, a way to capture the initial and final mental model that each student had of the circulatory system was developed. This analysis depicted the status of each student's mental model prior to and after learning, as a measure of representational changes occurring with deep understanding. In this case, status refers to the mental model's correctness and completeness in regard to the anatomical connections among the blood vessels, heart, lungs, and body.

The following section describes how a student's mental model was assessed. The student's initial model of how the circulatory system works as a whole was depicted from statements giving evidence of the anatomical connections and the direction of blood flow that students believed to exist in the system. These statements were taken only from the terms and blood path sections of the pretest (for the model prior to learning) and then again from these same sections of the posttest (for the model after learning). Unlike the question and answer sections of the pre- and posttests, which often challenged the students' initial conceptions of how the circulatory system works, both the terms and blood path sections allowed students to explain their knowledge of the system without interference from direct questions. It was for this reason that only the protocols from the terms and blood path sections were used for this analysis.

All relevant statements and evidence from the blood path drawings concerning the flow of blood and relations between circulatory components were identified and then combined to form a model of circulation. This involved integrating statements that were sometimes compatible and sometimes conflicting. These statements were graphically represented with drawings of connections between components and the direction of blood flow. As more statements were read, more and more detail could be added to the drawing. For the most part, students' protocols were rich with statements suggesting how the system worked. Often students would attempt to explain the whole system (at least the general workings of the system) in a few well-integrated

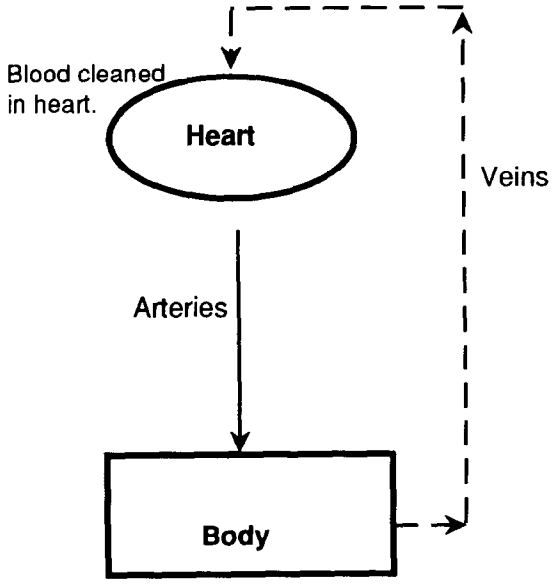

Figure 7. Depiction of a student's mental model.

statements. For example, one student explained the term "artery" in the pretest as follows:

> Artery is a general term for all tubes that are from the heart and they carry the clean blood from the heart to all the body. . . it [the body] always needs clean blood and the blood travels once through the arteries and when it's used, it travels back up in the veins to go back to the heart, the heart cleans it again, ummm, replenishes it with oxygen, umm and then it goes again to all the parts of the body.

From this explanation of one term the student was credited with the knowledge that:

1. Blood flows from the heart to the body in arteries.
2. Blood flows from the body to the heart in veins.
3. The body used the "clean" blood in some way, rendering it unclean.
4. Blood is "cleaned" or "replenished with oxygen" in the heart.
5. Circulation is a cycle.

These statements might be represented in a drawing as shown in Figure 7. This initial model was further enriched by considering and integrating additional statements from the protocols. For example, the same student later explained the term "heart":

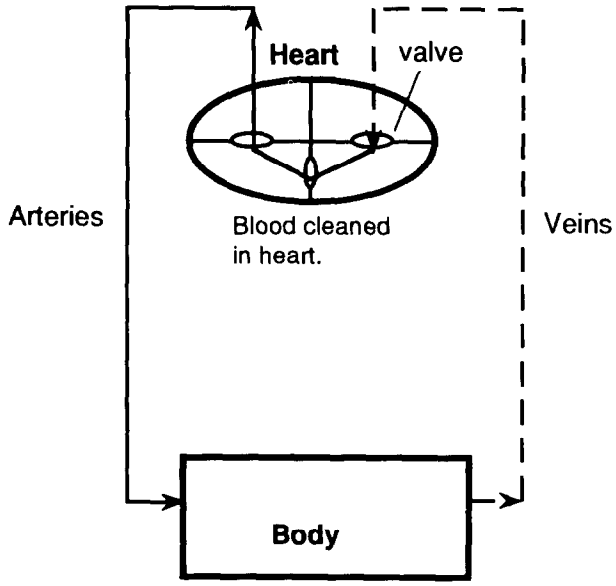

Figure 8. Continued depiction of a student's more enriched mental model.

The blood goes in at the upper right chamber and it then goes down to the downward right chamber, then it goes to the downward left chamber then it goes to the upward chamber then it goes out of the heart. And each chamber is divided by a valve that makes sure the blood goes in one direction.

This was integrated with the previous statements, adding more detail with the blood flow through the heart, resulting in a more enriched model, as shown in Figure 8.

A representation of each student's initial and final mental model was constructed in this manner, after several iterations through the protocols. At this level of analysis, an attempt was made to capture as much detail of each student's model as was possible. A great deal of variation existed between students' models, particularly in the functioning, behavior, and structure of specific components as well as relations between components. For example, this student's model is a single loop type of model where the blood flows from the heart to the body and then back to the heart with an added feature that the heart oxygenates the blood. She also talked about a specific blood flow through the heart involving valves and chambers. Another student had a similar single loop type of model, where the blood flowed from the heart

to the body and then back to the heart, but with the added feature of the lungs delivering air to the heart in order for oxygenation to occur there. And still another student with this single loop type of model did not even consider that blood was oxygenated anywhere in the system. These models are quite different in their details, yet they all have the basic overall structure of being a single loop: Blood flows from the heart to the body and then blood returns to the heart from the body. Six different general types of models were discerned, ranging from least accurate to most accurate: (1) no loop model; (2) ebb and flow model; (3) single loop model; (4) single loop with lungs model; (5) double loop—1 model; and (6) double loop—2 model. The double loop—2 model depicts the most accurate flow of blood through the circulatory system.

Because this first attempt at constructing models yielded various models in a more or less subjective way, this subjective analysis was validated by defining a set of necessary features for each of the six models, as shown in Table 4. Each student's model was then reclassified based on these features. Evidence for every feature had to be exhibited in order for that particular classification to be given to that model. For example, in order to be credited with a double loop—2 model, a student's protocols had to mention all the features listed for the double loop—1 model, plus all the ones listed for the double loop—2 model. Thus, the criteria for achieving the double loop—2 model were very stringent. This second pass through the verbal protocols, looking for specific features, served as a validity check for the first more subjective analysis.

Looking at all 24 subjects as a whole (both prompted and unprompted), there was an overall improvement in the formation of a correct mental model from reading the text (see Table 5), consistent with the overall gain in the question scores, as shown in Figure 2. The majority of students started with less accurate models such as the no loop model (21%) and the single loop model (50%), but after reading the text, the majority of students formed more accurate models, notably, double loop—1 (25%) and double loop—2 (46%). So reading the text and self-explaining did seem to improve the accuracy of students' mental models.

Prompted versus Unprompted Groups. Prompted students developed a correct model more often than unprompted students. Comparing the results of the posttest models, 8 of 14 (57%) prompted students attained the most accurate double loop—2 model, whereas only 2 of 9 unprompted students (22%) reached this level of accuracy in the formation of their models (see Table 5). (Student 19 in the unprompted group was excluded from this 22% calculation as having achieved the double loop—2 model because he began with a double loop—2 model.)

TABLE 4
Necessary Features for Each Type of Mental Model

No Loop
1. Blood is pumped from the heart to the body.
2. Blood does not return to the heart.

Ebb and Flow
1. Blood is primarily contained in blood vessels.
2. Blood is pumped from the heart to the body.
3. Blood returns to the heart by way of the same blood vessel.

Single Loop
1. Blood is primarily contained in blood vessels.
2. Blood is pumped from the heart to the body.
3. Blood returns to the heart from the body.

Single Loop with Lungs
1. Blood is primarily contained in blood vessels.
2. Heart pumps blood to body or to lungs.
3. Blood returns to heart from body or from lungs.
4. Blood flows from lungs to body or from body to lungs without return to heart in between.
5. Lungs play a role in the oxygenation of blood.

Double Loop—1
1. Blood is primarily contained in blood vessels.
2. Heart pumps blood to body.
3. Blood returns to heart from body.
4. Heart pumps blood to lungs.
5. Blood returns to heart from lungs.
6. Lungs play a role in the oxygenation of blood.

Double Loop—2
1. All features from Double Loop—1
2. Heart has four chambers
3. Septum divides heart lengthwise—sense of preventing mixing of blood.
4. Blood flow through heart is top to bottom.
5. At least three of the following:
   Blood flows from right ventricle to the lungs.
   Blood flows from lungs to left atrium.
   Blood flows from left ventricle to body.
   Blood flows from body to right atrium.

High Explainers versus Low Explainers. Within the prompted group, all 4 high explainers (100%) attained the most accurate double loop—2 model, whereas only 1 of 4 low explainers (25%) developed a model to this level (see Table 5 again). This shows that high explainers, those generating on average 87 explanations, all reached an understanding of the ultimate correct double-loop structure. Low explainers, those generating on average 29 explanations, attained the double loop—2 model in about the same proportion as the 9 unprompted students.

TABLE 5
Prompted and Unprompted Students' Pre- and Posttest Mental Models

Student	Inferences	Pretest Model	Posttest Model
Prompted			
High			
1	111	Single loop	Double loop—2
2	87	Single loop–lungs	Double loop—2
3	84	Single loop	Double loop—2
4	64	Single loop	Double loop—2
Medium			
5	47	Single loop	Single loop–lungs
6	47	Single loop	Double loop—2
7	42	Ebb and flow	Double loop—2
8	40	Single loop	Double loop—2
Low			
9	35	Single loop–lungs	Double loop—2
10	33	No loop	Single loop–lungs
11	28	Single loop	Double loop—1
12	20	Single loop–lungs	Double loop—1
Excluded			
13	20	Single loop	Single loop (no CAT score)
14	7	No loop	No loop (outlying CAT score)
Unprompted			
15		Single loop	Undeterminable
16		Single loop–lungs	Double loop—2
17		No loop	Double loop—1
18		Single loop	Double loop—1
19		Double loop—2	Double loop—2
20		Ebb and flow	Double loop—1
21		No loop	No loop
22		Single loop	Double loop—1
23		No loop	No loop
24		Single loop	Double loop—2

DISCUSSION

Eliciting self-explanations clearly enhances learning and understanding of a coherent body of new knowledge, whether one compares the amount learned by the prompted and the unprompted students, or whether one compares the amount learned by the high and low explainers. In either case, generating a large number of self-explanation inferences facilitated learning, when pre- to posttest gain scores were considered. Self-explaining also promoted

deeper understanding of expository materials, as displayed by the high explainers' ability to answer more complex questions, as well as being more able to induce the implicit function of a component. Besides assessing the display of understanding, understanding was also captured by changes in the students' mental models of the circulatory system. Many more prompted students, including all the high explainers, represented the circulatory system by the correct double loop—2 model, whereas few unprompted students achieved the correct model. Clearly, not only does eliciting self-explanations promote greater learning, but the more students self-explain, the deeper their understanding. Therefore, learning a body of declarative knowledge can neither simply be the mere instantiation of existing stored knowledge nor the direct encoding of it. Rather, learning is the use of existing knowledge in conjunction with new information to create more new knowledge. In some ways, self-explaining is thinking with what one knows.

What might mediate learning from self-explaining? Three processing characteristics of self-explaining may shed light on why it is a particularly effective learning activity. The first characteristic of self-explaining is that it is a constructive activity. This means that new declarative or procedural knowledge is constructed. Constructed declarative knowledge was directly captured from the explanation protocols in this study as explanation inferences, and constructed procedural knowledge was captured in the Chi et al. (1989) physics explanation protocols as *if–then* conditional rules (Chi & VanLehn, 1991).

Constructed procedural rules can be captured in other indirect ways as well, namely, by developing a rule-based model for solving problems, and then seeing when these rules are used. VanLehn, Jones, and Chi (1992) found that only 50–60% of the rules needed to solve physics problems of the kind used in Chi et al. (1989) were presented in the physics text. This means that the remaining rules needed for solving problems had to be constructed by the solver. Presumably, self-explaining facilitated the construction of these rules. This finding reiterates the point made in the introduction, that learning a procedural skill cannot be solely the direct encoding and subsequent speeding up of instruction.

An alternative way to infer the construction of rules is to examine the problem-solving protocols in the Chi et al. (1989) data for errors committed by high explainers (those who generated more than 40 self-explanations) and low explainers (those who generated fewer than 20 self-explanations across the three examples). Missing rules was the only category where high explainers made significantly fewer errors than low explainers (VanLehn & Jones, 1993). Low explainers seemed to have more errors arising from the lack of usable rules, which presumably could have been constructed had they self-explained more. Thus, in the physics study, self-explanations seem to provide opportunities to construct rules that can subsequently be used during problem solving; in the current biology study, self-explanations pro-

vide opportunities to construct knowledge inferences (such as inducing the function of a component) that can be subsequently used to answer complex questions.

The fact that self-explaining is a form of constructive activity brings into question whether other forms of nonverbal constructive activity are equally effective at enhancing learning because the majority of the ones investigated in the literature (such as peer problem solving and reciprocal teaching) are verbal in nature. This issue can be partially addressed by examining the diagram-drawing activity that accompanies self-explanations. Although high explainers drew more diagrams than low explainers, there was also a greater variability among high explainers. Low explainers, however, consistently drew fewer diagrams. The diagrams drawn were also very sketchy, hence they could not be coded for quality. These data do not address whether drawing diagrams per se, without self-explaining, would be an adequate form of constructive activity to promote learning. Basically, to the extent that an activity (such as diagram drawing) is a constructive one, then it could probably facilitate learning to some degree.

The process of self-explaining contains an important second characteristic: It encourages the integration of newly learned materials with existing knowledge. Taken at face value, people often wonder how students can construct explanations when they have incomplete knowledge of the domain. The data here show that at least 30% of the self-explanations are produced from integrating new information with old knowledge. The implication of this characteristic of self-explaining is that other kinds of constructive activities that discourage integration, such as summarizing, may be less effective at promoting learning.

An outcome of integration with prior knowledge is that a self-explanation can result in an incorrect piece of knowledge. In both the physics data and the current data, about one fourth of the self-explanations are incorrect, although generating incorrect self-explanations was not detrimental to learning (because the pattern of results remained the same when incorrect ones were partialed out). It is even conceivable that generating incorrect self-explanations can provide a learning experience. Here is one possible interpretation. Having articulated an incorrect explanation, a student continues to read the next sentence or sequence of sentences in the text. Eventually, the text sentences, because they always present correct information, may contradict knowledge embodied in the incorrect self-explanation. This will create a case of conflict. Hence, one interpretation is that creating an incorrect self-explanation merely objectifies that piece of knowledge, which allows it to be examined in the face of conflicting information from subsequent sentences, thus establishing the opportunity for self-repair to resolve the conflict. Suggestions that much of learning takes place during confrontations, conflict situations, or at points of impasses have been presented in the developmental (Kuhn, 1972), social (Doise, Mugny, & Perret-Clermont,

1975), and cognitive science literature (VanLehn, 1988). Presumably, such confrontations and conflicts create opportunities for resolving them.

However, the fact that self-explanation involves the integration of new knowledge with old knowledge suggests that learning a new body of knowledge must also take into account the nature and coherence of students' existing mental models and/or conceptions. For some concepts (such as forces, heat, natural selection), the nature of the initial (mis)conceptions are "ontologically incompatible" with the veridical conceptions. That is, the students' initial conceptions of these concepts belong to a different ontological category, so that integrating new information with them leads to further misunderstandings that are robust and persistent (Chi, 1992, 1993; Chi & Slotta, 1993; Chi, Slotta, & de Leeuw, 1994). It is unclear at this point whether self-explaining can facilitate understanding of these ontologically incompatible concepts. In the case of ontologically compatible concepts and systems (such as the circulatory system), however, integration of new information with existing faulty conceptions allows the new information to revise the initial false beliefs more readily (de Leeuw, 1993; an example of this revision will be shown shortly). The remaining discussion applies only to ontologically compatible concepts.

Although an incorrect self-explanation or a faulty initial mental model can produce conflicts and confrontations, it is not always the case that conflicts and contradictions lead to resolution and learning, as some of the literature on the role of anomalies shows (see discussion in Chi, 1992, pp. 152–158). However, a plausible account of why self-explaining may be a powerful mechanism for removing conflicts has to do with its third processing characteristic, namely, that self-explaining is carried out in a continous, ongoing, and piecemeal fashion, thereby often resulting in partial, incomplete, and fragmented self-explanations. Suppose one thinks of self-explaining as the process of creating or revising a mental structure (in this biology case, a mental model of the system, and in the previous physics case, a mental structure of the solution example). While reading (either expository text sentences or example statements), there are many opportunities whereby what is read contradicts what is being created or existed a priori in one's mental structure. Self-explaining thereby gives rise to multiple opportunities to see conflicts between one's evolving mental structure and the veridical description of it from the text. For instance, Student 9's initial mental model was a single loop with lungs that links the blood path from the lungs directly to the rest of the body, without another stopover in the heart. He said during the pretest of the term "blood" that

> blood from the heart into the lungs, where oxygen is taken out of the lungs and held in the blood and then, uh. . . *the blood goes throughout the rest of the body*.

And then, again, during the blood path portion of the pretest, he said

> [blood] *comes back out through the lungs and then goes all over the body* through arteries.

Upon reading Sentences 33 and 34 in the text as shown

> Sentence 33: In the lungs, carbon dioxide leaves the circulating blood and oxygen enters it.
> Sentence 34: The oxygenated blood returns to the left atrium of the heart.

he then explained:

> Okay. It's just saying that the blood has to *go back to the heart before it goes to the rest of the body*. So it's pumped through another time.

Thus, the explanation generated upon reading Sentence 34 obviously revised his initial mental model, which omitted the blood's stopover at the heart before going on to the rest of the body. The point of this example is that ongoing piecemeal self-explaining allows conflicts to be recognized and resolved at many loci, where the changes are more minute and more easily repaired.

Contrast this with an alternative case, in which a student read (probably without self-explaining) a theory and attempted a complete explanation using that theory. An excellent illustration of this outcome can be seen in some protocols that Scardamalia (1992, July) collected. After reading about blood clots, a student was asked to explain how a cut heals. The first response the student gave, on the basis of reading the text material, was the following:

> When you get a cut it bleeds. In your blood there are things called platelets. Platelets made a shield on your cut, it is called a scab, it protects your skin when it is healing.

The student was then asked to use his own theory to answer the same question. Here is the student's second response to the same question:

> My theory is that when you get a cut the blood vessel that got cut dies and the heart stops sending blood to that vessel until it heals.

What is so telling about these two explanations is that each is complete, and each is generated from a coherent theory, and each theory has its own components (platelets and scab in the scientific explanation, and dying blood

vessel and heart stops sending blood in the naive explanation). That is, the student's old theory obviously was not subverted by the new theory *when* the new theory was learned, as was the case in the example of Student 9's revision of his old theory about the destination of blood path after leaving the lungs. Although the student may see that one explanation is pitted against the other, it is difficult to see how the two opposing explanations can be integrated and resolved at this global complete explanation level. Noticing conflicts and resolving them must be attempted at a more fine-grained piecemeal level, the level of what happens to blood when one gets a cut (whether it bleeds, or whether the heart stops sending blood). Thus, explaining on the basis of the text's theory in an isolated and self-contained way cannot be taken to mean that the student understands it and has rejected other explanations. In order to assimilate the new theory exclusively, it must be incorporated with existing knowledge at a fine-grained level. This example and our data emphasize the importance of the second and third processing characteristics of self-explaining, namely, that learning involves the *integration* of new knowledge with old knowledge, and such integration is best carried out in a more *minute* and *ongoing* fashion.

These second and third processing characteristics have direct implication for what kind of constructive activity might be more or less efficient at promoting learning. Some existing training studies have requested that students generate explanations that are scientific (Coleman, 1992), and/or based on the scientific theory presented in the text (Ohlsson, 1992). These kinds of constructive activities may preclude and discourage students from integrating their prior misconceived knowledge with the newly acquired scientific knowledge, thereby possibly producing isolated and self-contained explanations, as the preceding quotation showed about what happens to a cut. Instead of asking students to generate complete explanations based on a given theory, it might be more productive to ask them to generate microscopic and partial self-explanations because these minute and fragmented ones may have a better chance of repairing one's erroneous initial mental model. In sum, self-explanation has three important characteristics that may contribute to its effectiveness as a learning skill. Many alternative activities, such as diagram drawing, summarizing, generating complete theory-based explanations, may be more limited as *learning skills* because each of them lacks all of the characteristics of self-explanations. However, these alternative activities may be ideal for other purposes, such as means of assessing understanding.

The general claim of this article, that generating more self-explanations promotes greater learning, is supported by the contrast between the prompted and unprompted conditions, as well as between the high and low explainers. However, the fact that some prompted students generated more self-explanations than others, suggests that perhaps some other underlying factors may be mediating the effectiveness of prompting. The results here sug-

gest at least two possible alternative factors. First, although the benefit of self-explaining seems to be independent of ability (or achievement) and prior domain-relevant knowledge (as assessed in the pretests) it is possible that prior *general world* knowledge may play a key role because at least 30% of the self-explanations resulted from integrations with prior commonsense knowledge. A second alternative mediating factor may have to do with the strategy of integrating the new to-be-learned materials with previously encoded text information in a more extensive way, namely, across topics rather than just within a given topic (because the high explainers had a much higher proportion of across-topic links). This may be a strategy that can be taught in order to achieve maximal generation of explanation inferences. In this study, because we had visions of implementing an automated prompting system, we wanted to see the extent to which a very nonspecific and unguided prompt to elicit self-explanations would promote learning. The results of this study certainly point to ways in which more extensive training can perhaps guide the elicitation of even more self-explanations. Successful efforts along the lines of using more elaborate training procedures can be found in Bielaczyc et al. (in press) and King (1994).

REFERENCES

Beck, I.L., McKeown, M.G., Sinatra, G.M., & Loxterman, J.A. (1991). Revising social studies text from a text-processing perspective: Evidence of improved comprehensibility. *Reading Research Quarterly, 26*, 251–276.

Bereiter, C., & Scardamalia, M. (1989). Intentional learning as a goal of instruction. In L.B. Resnick (Ed.), *Knowing, learning and instruction: Essays in honor of Robert Glaser*. Hillsdale, NJ: Erlbaum.

Bielaczyc, K., Pirolli, P., & Brown, A.L. (in press). Strategy training in self-explanation and self-regulation strategies for learning computer programming. *Cognition and Instruction*.

Chi, M.T.H. (1992). Conceptual change within and across ontological categories: Examples from learning and discovery in science. In R. Giere (Ed.), *Cognitive models of science: Minnesota Studies in the philosophy of science*. Minneapolis: University of Minnesota Press.

Chi, M.T.H. (1993, June). Barriers to conceptual change in learning science concepts: A theoretical conjecture. *Proceedings of the 15th Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.

Chi, M.T.H. (1994). Analyzing the content of verbal data to represent knowledge: A practical guide. Submitted to *Journal of the Learning Sciences*.

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

Chi, M.T.H., & Slotta, J.D. (1993). The ontological coherence of intuitive physics. Commentary on A. diSessa's "Toward an epistomology of physics." *Cognition and Instruction, 10(2 & 3)*, 249–260.

Chi, M.T.H., Slotta, J.D., & de Leeuw, N. (1994). From things to processes: A theory of conceptual change for learning science concepts. *Learning and Instruction, 4*, 27–43.

Chi, M.T.H., & VanLehn, K. (1991). The content of physics self-explanations. *Journal of the Learning Science, 1*, 69-105.

Coleman, E. (1992, April). *Reflection through explanation*. Paper presented at the American Educational Research Association meeting, San Francisco.

deKleer, J., & Brown, J.S. (1983). Assumptions and ambiguities. In D. Gentner & A.L. Stevens (Eds.), *Mental models*. Hillsdale, NJ: Erlbaum.

de Leeuw, N. (1993, June). Students' beliefs about the circulatory system: Are misconceptions universal? *Proceedings of the 15th Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.

Doise, W., Mugny, G., & Perret-Clermont, A.N. (1975). Social interaction and the development of cognitive operations. *European Journal of Social Psychology, 5*, 367-383.

Dole, J.A., Valencia, S.W., Greer, E.A., & Wardrop, J.L. (1991). Effects of two types of prereading instruction on the comprehension of narrative and expository text. *Reading Research Quarterly, 26*, 142-159.

Ericsson, K.A., & Simon, H.A. (1984). *Protocol analysis: Verbal reports as data*. Cambridge, MA: MIT Press.

Ferguson-Hessler, M.G.M., & de Jong, T. (1990). Studying physics texts: Differences in study processes between good and poor performers. *Cognition and Instruction, 7*, 41-54.

Goldman, S.R., & Duran, R.P. (1988). Answering questions from oceanography texts: Learner task and text characteristics. *Discourse Processes, 11*, 373-412.

Graesser, A.C., & Murachver, T. (1985). Symbolic procedures of question answering. In A.C. Graesser & J.B. Black (Eds.), *The psychology of questions*. Hillsdale, NJ: Erlbaum.

Haugeland, J. (1978). The nature and plausibility of cognitivism. *Bahavioral and Brain Sciences, 2*, 215-260.

Hawkins, J., & Pea, R.D. (1987). Tools for bridging the culture of everyday and scientific thinking. *Journal for Research in Science Teaching, 24*, 291-307.

King, A. (1994). Guiding knowledge construction in the classroom: Effects of teaching children how to question and how to explain. *American Educational Research Journal, 31*, 338-368.

Klahr, D., & Dunbar, K. (1988). Dual space search during scientific reasoning. *Cognitive Science, 12*, 1-48.

Kuhn, D. (1972). Mechanisms of change in the development of cognitive structures. *Child Development, 43*, 833-842.

Lemke, J.L. (1990). *Talking science: Language, learning and values:* Norwood, NJ: Ablex.

Marton, F., & Saljo, R. (1976). On qualitative differences in learning. *British Journal of Educational Psychology, 46*, 4-11.

Mayer, R.E. (1993). Illustrations that instruct. In R. Glaser (Ed.), *Advances in instructional psychology* (Vol. 4). Hillsdale, NJ: Erlbaum.

Nathan, M.J. Mertz, K., & Ryan, B. (1993). *Learning through self-explanation of mathematical examples: Effects of cognitive load*. Paper presented at the 1994 Annual Meeting of the American Educational Research Association.

Ng, E., & Bereiter, C. (1991). Three levels of goal orientation in learning. *The Journal of the Learning Sciences, 1*, 243-272.

Nicholas, D.W., & Trabasso, T. (1980). Toward a taxonomy of inferences for story comprehension. In F. Wilkening, J. Becker, & T. Trabasso (Eds.), *Information integration by children*. Hillsdale, NJ: Erlbaum.

Nix, D. (1985). Notes on the efficacy of questioning. In A.C. Graesser & J.B. Black (Eds.), *The psychology of questions*. Hillsdale, NJ: Erlbaum.

Ohlsson, S. (1992). The cognitive skill of theory articulation: A neglected aspect of science education? *Science & Education, 1*, 181-192.

Pirolli, P.L. & Recker, M. (1994). Learning strategies and transfer in the domain of programming. *Cognition and Instruction, 12*, 235-275.

Russell, D.M., & Kelley, L. (1991, April). *Using IDE in instructional design: Encouraging reflective instruction design through automated design tools*. Paper presented at Annual Conference of American Educational Research Association.

Scardamalia, M. (1992, July). *The role of adaptation and understanding in knowledge-building communities*. Paper presented at conference of NATO Advanced Science Institute, Kolymbari Crete, Greece.

Slamecka, N.J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory, 4*, 592–604.

Towle, A. (1989). *Modern biology*. New York: Holt, Rinehart & Winston.

VanLehn, K. (1988) Student modeling. In M. Polson & J. Richardson (Eds.), *Foundations of intelligent tutoring systems*. Hillsdale, NJ: Erlbaum.

VanLehn, K., & Jones, R.M. (1993). What mediates the self-explanation effect? Knowledge gaps, schemas or analogies? In M. Polson (Ed.) *Proceedings of the 15th Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.

VanLehn, K., Jones, R.M., & Chi. M.T.H. (1992). A model of the self-explanation effect. *Journal of the Learning Sciences, 2*, 1–59.

Webb, N.M. (1989). Peer interaction and learning in small groups. *International Journal of Education Research, 13*, 21–39.

Zajonc, R.B. (1960). The process of cognitive tuning in communication. *Journal of Abnormal and Social Psychology, 61*, 159–167.

APPENDIX

Instructions Given to Prompted Subjects Prior to Reading the Text

The following is a chapter on the human circulatory system which was taken from a high school text book. We are trying to learn more about how students read and learn from a textbook, as well as what makes some textbooks better than others. In order for us to assess what information the text book is good at making understandable, it is important that you read every line very carefully—as if studying for an exam.

The text is presented one sentence at a time so that you will have time to really think about what information each sentence provides and how this relates to what you've already read.

We would like you to read each sentence out loud and then explain what it means to you. That is, what new information does each line provide for you, how does it relate to what you've already read, does it give you a new insight into your understanding of how the circulatory system works, or does it raise a question in your mind. Tell us whatever is going through your mind—even if it seems unimportant.

You may need to go back and re-read parts of the text to really understand all the material. Also, some people find it helpful, when reading difficult material, to draw a picture or take notes. Please feel free to do what is best for you—please use these transparencies for this purpose. Let me know when you'd like to start a new transparency.