Can a Computer Interface Support Self-explaining?

Robert G.M. Hausmann, M.S.
Michelene T.H. Chi, Ph.D.
University of Pittsburgh

Previous research has shown that when an experimenter or a tutor prompts students to self-explain orally, generating such self-explanations is effective for learning. If self-explanations are readily produced by prompting, then it would be trivial to implement an automated prompting system using a computer interface. In an attempt to replicate previous research using a human prompter with spoken self-explanations, two experiments were designed using a computer prompter with typed self-explanations. The first experiment tested the effectiveness of spontaneously typed self-explaining while using a computer interface without prompting. The results showed that the amount of self-explaining was surprisingly low, given the amount observed in past research. Typing seems to have caused the students to paraphrase the materials instead. The second experiment tested the effectiveness of an automatic computer prompter, as compared to a human prompter using the same interface. Automatic prompting was just as effective as human prompting, and prompting did increase the amount of typed self-explanations and learning.

The construction of explanations to oneself, while studying a worked-out example problem (Chi & Bassok, 1989; Chi, Bassok, Lewis, Reimann, & Glaser, 1989; Chi & VanLehn, 1991) or while studying a text, is a domain-general learning strategy (Chi, DeLeeuw, Chiu, & LaVancher, 1994). Chi et al. (1989) demonstrated that the successful problem solvers spontaneously generated more self-explanations when studying the domain material (e.g., physics) than the less successful students. Furthermore, the successful students generated more self-explanations that could be inferred from Newton's three laws, whereas the less successful students generated very few principle-based explanations. These findings suggest that successful students, when attempting to learn procedural knowledge, spontaneously generated more and deeper explanations to themselves, even though they came into the learning situation with equivalent prior knowledge as the less successful students. The beneficial outcome of the process of constructing explanations, with the goal of increasing one's understanding, is called the self-explanation effect (Chi, 2000).

Moreover, students can be prompted to generate more self-explanations than they might produce spontaneously. Chi et al. (1994) contrasted the learning of students who were prompted by an experimenter against the learning of a control group. The prompted students gained more in terms of their pre-test to post-test difference. Furthermore, the difference in learning between the two groups was more pronounced for the more difficult questions. Restricting analysis within the prompting group, an analysis of Good and Poor students suggested they both increased at the same rate (32% versus 30% respectively).

Computer Interface and Self-explanation

Because elicitation by a human prompter can be expensive, one obvious application of self-explaining is a computer system that prompts the student to self-explain while learning the domain material (Chi et al., 1994). Research that explores the effectiveness of computer support for self-explanation have generally found an increase in learning (Aleven & Koedinger, 2000; Aleven, Koedinger, & Cross, 1999; Howie & Vicente, 1998; Renkl, 1997). For example, Aleven et al. (1999) found a positive learning gain for students using their PACT geometry tutor. The participants were required to select their reasons for each step of the problem solving solution from a glossary (or menu) of geometry rules and definitions. Aleven et al. (1999) found the students who were required to provide a reason for their solutions to the geometry problems did significantly better on post-test than students in the control condition. Occasionally, some studies have found mixed results (Conati & VanLehn, 1999, 2000). For example, Conati and VanLehn (1999, 2000), using a similar menu-selection method of self-explaining, found such prompting to be helpful only for some population of students, but not others. We will hypothesize in Experiment 1 why menu-selection is not always an effective means of generating self-explanations.

Instead of learning in a procedural domain (e.g., physics or geometry), the present study investigated the effects of self-explaining in a conceptual domain: the circulatory system. The students were asked to read a passage that covers the path of blood as it circulates throughout the human body. Components of the heart, such as the various chambers and valves, are described in terms of their structure, function, and location. The overall goal for the student was to develop a mental model that approximates the scientifically accepted model. For instance, some believe the heart pumps the blood out to the body and back, without making any reference to pulmonary circulation, nor differentiating the chambers within the heart. This naïve "Single-Loop" model is contrasted with the scientifically accurate "Double-Loop" model in which the blood receives its oxygen from the lungs.

To facilitate comparisons with earlier studies (Chi et al., 1994; Chi, Siler, Jeong, Yamauchi, & Hausmann, 2001; Jeong, 1998; Wathen, 1997), the learning and assessment materials were taken from Chi et al. (1994). A few important differences exist between the present study and the original study by Chi et al. (1994). First, the materials used in the original study were designed for eighth graders, whereas the present study used college undergraduates. However, it has been shown that these materials are challenging even to college students, since some of them do maintain incorrect mental models after reading the text used in the present study (Jeong, 1998). Second, in an effort to include only the textual material relevant to circulation, the materials used for the present study were a subset of Chi et al. (1994). Details of the blood and other organs were excluded. Finally, the present study selected a representative subset of the original assessment questions.

Experiment 1

One possible reason why some populations do not benefit from menu-drive self-explanation is that choosing from a menu of explanations is not constructive. Those students who found the menu selection systems to be beneficial might be covertly self-explaining anyway regardless of its implementation (menu selection, spoken, typed, or otherwise). Therefore, the goal of the first experiment is to test the hypothesis that allowing students to be more generative (i.e., free-form typing) might allow for greater self-explaining and thereby learning, than selection from a menu (used in these other studies). The first experiment attempted to replicate the results from Chi et al. (1989), to measure the degree to which students will spontaneously generate explanations via the keyboard, while using a computer interface.

It is important to replicate the self-explanation effect in the typed modality because it is not clear what the effect of typing on self-explaining will be. Several relevant differences exist between speaking and typing. First, less effort is required to speak one's thoughts than to type a thought (Lebie, Rhoades, & McGrath, 1996). Second, one might be less embarrassed about saying incomplete or incoherent ideas because spoken speech does not leave an enduring trace (Clark, 1999). Finally, the ease of speaking allows ideas to flow quickly and naturally without being monitored, whereas one might filter many ideas before they are typed. This lack of filtering might explain why spoken self-explanations tend to be fragmented and incomplete (Chi, 2000).

Design

Conditions

Two conditions were contrasted. The spontaneous self-explanation condition served as the experimental condition, while the read-only condition provided a baseline measure of learning.

Spontaneous self-explanation. Students in the spontaneous self-explanation condition were asked to read the material and try to understand the circulatory system as best as possible. They were encouraged to "type comments to themselves," which was analogous to generating self-explanations. There were no time constraints on this control condition. The time spent reading the text and typing comments was measured.

Read-only. The control group, the read-only condition, was not given the opportunity to type any comments during the learning phase of the experiment. Instead, the read-only condition was presented with one sentence for a fixed amount of time. The presentation time for each individual sentence was the average reading time for each corresponding sentence in the self-explanation condition. The average reading time from the self-explanation condition was used to control for time-on-task. The reason for including this control condition was to replicate the self-explanation effect (Chi et al., 1989), while controlling for time-on-task (Renkl, 1997). If a gain in learning is observed for the students who self-explain, over and above the students who merely read the text on their own, then the results will provide evidence that self-explaining in a computerized environment can be effective.

Method

Participants

Each condition included 10 students, for an overall total of 20 participants. All participants were selected from a pool of undergraduate psychology majors, and they were given course credit upon completion of the experiment. Because the study required the participants to type during the pre-test, the learning phase, and the post-test, students who were selected to participate were average typists. However, no formal evaluation of typing speed was assessed.

Materials

Text. The text consisted of a passage of 62 sentences, taken from a popular biology textbook chapter on the human circulatory system (Towle, 1989). Each sentence was presented individually via a 17-inch, color computer monitor. Students' typed statements were captured and stored in a spreadsheet, and were later segmented and coded.

Assessment materials. The pre-test consisted of two parts: the blood-path drawing was administered first and vocabulary test followed. The blood-path drawing served to capture the student's mental model of the circulatory system. It consisted of drawing, while explaining aloud, where and how the blood circulates throughout the body. A vocabulary test required the participant to define 21 terms relating to the circulatory system.

To assess the degree to which each group learned during the experiment, the same blood-path and vocabulary test were re-administered for the post-test. In addition to these two tasks, a set of questions was used to assess the knowledge gained by the students during the tutoring sessions. The questions were designed to reflect a range of difficulty (see Figure 2, p. 458 from Chi et al., 1994 for evidence of question difficulty). The easiest set of questions (Category 1) is considered "verbatim questions." The answers to these questions were often explicitly stated in one of the sentences in the text. Category 2 questions were more difficult because the reader was required to make some inferences between sentences of the text. Category 3 questions were much more difficult because the students were not only required to generate inferences from the text, but they also needed to integrate their background knowledge with their inferences. Category 4 questions were the most difficult, requiring the students to generalize their knowledge to health related issues concerning the cardiovascular system.

Procedure

Pre-test

Before beginning the experimental session, students provided the experimenter with self-report SAT-verbal scores. Then the students were administered a pre-test, which consisted of the blood-path drawing and the vocabulary test. For the blood-path drawing, the students were given an outline of the human body and asked to explain how the blood travels throughout the body. Their explanations were tape recorded and later analyzed. Upon completion of the blood-path drawing, the students moved to the computer where they were administered the vocabulary test. Each term was presented one at a time, and the participants were required to type their definitions into a field located at the bottom of the computer interface. After completing the pre-test, the students began the learning phase of the study.

Learning Phase

Students from both conditions read the material, which was presented on a computer monitor. As stated previously, different groups were told to read or type comments. Both conditions had the same interface and read the same text. The interface consisted of a field at the top of the screen where the text appeared. At the bottom of the screen, labeled "Student Message," contained a field where the student could type his or her comments. When finished typing, the student clicked the "Submit" button for the next sentence from the text. When the learning phase was over, the computer prompted the student to alert the experimenter.

Post-test

After the learning session, a three-part post-test was administered to measure the students' knowledge: the blood-path drawing, with the vocabulary test and the Category 1–4 questions presented via computer.

Results

Scoring the Tests and Mental Models

For the vocabulary test, each of the 21 vocabulary words was scored on a 3-point scale. A definition was given a "2" if the answer was complete; a "1" was assigned to a definition that was partially correct, but left out a significant relationship or structure; and a "0" was provided for any answer that was factually wrong or left blank. The Category 1–4 questions were graded on a 1- or 2-point scale. If the answer only required one idea or relationship, then one point was assigned. If the answer had two parts, then two points were assigned.

Students tend to represent the circulatory system according to one of seven different versions of the scientifically correct model of the circulatory system. Mental model categories six and seven approximate the correct Double-Loop model, with minor variations between them. Details of how mental models are captured from protocols are described in Chi et al. (1994). Briefly, the characteristics of each mental model are searched for in the protocol. For example, if the student mentioned that the heart provides blood to the lungs, then there is evidence that the student understands that the lungs are involved in circulation (i.e., a mental model of four or higher). Further evidence is sought to ascertain whether or not the student understands that the lungs are for the oxygenation of blood. After a student's mental model is captured, each mental model was categorized. For instance, a student's model was characterized as category six, or also called the Double-Loop (1) model, if all the constituent pieces of the heart were mentioned, but in the wrong locations. If the model matched exactly to the scientifically correct model, then the student's model was scored as a seven, or Double-Loop (2) model.

Equivalent Groups

Overall, the spontaneous self-explanation and the read-only groups were equal in terms of ability and pre-test knowledge. There were no reliable differences between groups in terms of their self-reported SAT-verbal scores ($t(18) = 0.42$, $p = 0.68$), pre-test vocabulary ($t(18) = 0.002$, $p = 0.97$), as well as pre-test mental model ($p = 0.56$, Wilcoxon Two-sided Test). Therefore, because the groups were equivalent in pre-test knowledge and ability, the remaining analyses were conducted without statistically controlling for these variables.

Pre-test to Post-test Learning

Three measures were used to measure learning: the vocabulary test, the mental model analysis, and the Category 1–4 questions. For the vocabulary test, both groups significantly improved their understanding of the terms that describe the circulatory system (see Table 1). Both groups, however, gained about the same amount (around 24%) of vocabulary, which was a significant increase within each condition ($p < 0.0001$ for each group), but not between conditions.

Table 1

Pre-test and Post-test Performance on the Vocabulary Test for the Spontaneous Self-Explanation (SSE) and Read-Only (RO) Conditions

		Pre-test		Post-test		
Condition	*n*	*M*	*SD*	*M*	*SD*	*p value for paired t-test*
SSE	10	48.1%	13.4	72.9%	12.2	0.0001
RO	10	48.3%	13.2	71.7%	9.6	0.0001

Note. There were no reliable differences between the two groups for either the pre- or post-test.

For the mental model analysis, both conditions started with relatively similar pre-conceptualization of the circulatory system. That is, the distribution of the types of mental models of the students in each group was roughly the same, in that 60% of the students in both groups started out with a Single-Loop model, consistent with the results reported in Chi (2000) (see Table 2). After reading the text about the circulatory system, both the read-only group and the spontaneous self-explanation group improved their mental model scores (80% to 60% respectively finished with a Double-Loop (1) model).

Table 2

Percentage of Mental Models During Pre-test and Post-test for the Spontaneous Self-Explanation (SSE) and Read-Only (RO) Conditions

Mental Model Category		Pre-test		Post-test	
Name	#	RO	SSE	RO	SSE
No Loop	1				
Ebb and Flow	2	20%	10%		
Single-Loop	3	60%	60%		20%
Single-Loop w/ Lungs	4				
Multiple Loops	5	10%	10%		
Double-Loop (1)	6	10%	20%	80%	60%
Double-Loop (2)	7			20%	20%

Note. Definitions for the mental models can be found in Table 4 of Chi et al. (1989). Double-Loops (1) and (2) are the most scientifically accurate models of the circulatory system.

To provide a more sensitive test of learning, Verbatim (Category 1) questions were differentiated from

Integration (Category 2–4) questions. The mean scores for the Verbatim (Category 1) and Integration (Category 2–4) questions are shown in Table 3, and subject to an analysis of variance (ANOVA), with Condition (read-only and spontaneous self-explanation) as a between groups variable and Question (Verbatim and Integration) as a within-groups variable. Although previous literature suggests that self-explanation is beneficial for learning, there was again no evidence that the students in the spontaneous self-explanation condition learned any better than the read-only condition. If anything, the read-only group performed (slightly, but not reliably) better on the more difficult questions than the spontaneous self-explanation group, which parallels the mental model findings.

Table 3
Percentage of Correct Answers on the Verbatim (Category 1) and Integration (Category 2-4) Post-test Questions for the Spontaneous Self-Explanation (SSE) and Read-Only (RO) Conditions

		Verbatim Questions		Integration Questions	
Condition	*n*	*M*	*SD*	*M*	*SD*
SSE	10	61.7%	13.7	31.2%	10.7
RO	10	60.0%	17.9	38.4%	11.7
p value for main effects		0.82		0.17	

Note. The interaction between Question (Verbatim and Integration) and Condition (SSE and RO) was not significant.

Frequency of Self-explanation

Because previous literature predicted a benefit for self-explanation, the lack of difference between the two groups was surprising. To see whether students did in fact self-explain, the typed protocols were analyzed according to the coding scheme outlined in Chi et al. (1994). The protocols were analyzed for the following contents: self-explanations, paraphrases, and meta-cognitive statements. Self-explanations are statements of inference, which went beyond the information explicitly stated in the text. Paraphrases were defined as a verbatim summary of the currently presented sentence. Meta-cognitive statements were utterances that commented on the state of the learner's knowledge or understanding.

The frequency data for spontaneous self-explanation condition are presented in Figure 1. The data indicate that there were very few self-explanations during the learning phase. On average, about 1.0 self-explanation was generated across the 62 sentences. Instead of self-explaining, it seems that the students were more likely to paraphrase the material. Students typed, on average, 24.4 paraphrasing statements. Because paraphrasing dominated the content of the protocol, it was the only variable that positively correlated with learning, but only with Verbatim questions ($r = 0.68, p = 0.03$).

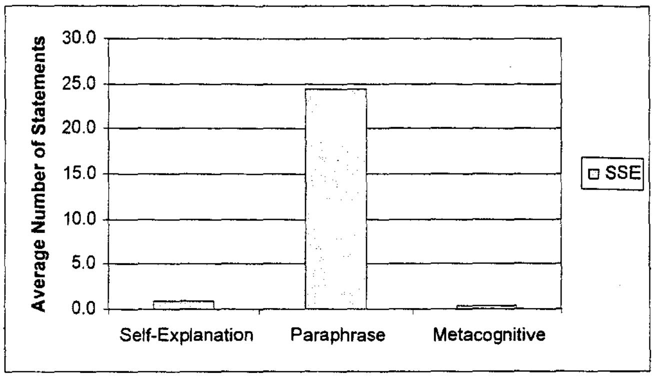

***Figure 1.* The average number of statements made by participants in the spontaneous self-explanation (SSE) condition during the learning phase of the study.**

Discussion

Prior studies on the self-explanation effect have traditionally asked students to self-explain orally. Students who self-explained the most often do better on later post-test measures. Because students generated very few self-explanations, the present study did not obtain the self-explanation effect. Two reasons might account for the lack of effect. First, the qualitative analysis revealed a lack of explicit self-explaining. It is therefore not surprising that the spontaneous self-explanation and the read-only groups showed similar learning outcomes because the students in the self-explanation condition did not engage in any constructive activities. Instead, they predominantly paraphrased, and such paraphrasing correlated only with verbatim knowledge. This correlation further confirms the assumption that paraphrasing only enables students to learn shallow knowledge, and this is why paraphrases originally were not considered to be self-explanations (Chi et al., 1989).

The second possible reason for the lack of the self-explanation effect is a difference in the modality by which the students self-explained. Past research asked participants to self-explain orally, while the present study asked students to type their comments. Generating comments in the typewritten format seems to inhibit self-explaining, for the various reasons entertained above. For example, in typing, students might have filtered out what they would spontaneously say orally. Confirmation for this interpretation can be noted in the completeness and accuracy of their paraphrases, whereas oral self-explanations typically tend to be fragmented and incoherent. Another possible reason that typing can inhibit self-explaining is that it uses more cognitive capacity (Gentner, 1988). In fact, the students who typed while learning did worse (although not significantly worse) than the students who were asked merely to read the text. Having to type may have reduced their cognitive resources for learning.

Because very little self-explaining was observed in the spontaneous self-explanation condition, a second experiment was conducted to overcome the limitations of keyboarding by prompting the students to self-explain.

Experiment 2

While self-explaining is particularly effective for good students, not all students engage in constructive activities when left to their own devices. Chi et al. (1994) provided evidence for the benefit of eliciting students to engage in self-explaining using content-free prompts. Using a conceptual domain, they found that students who were prompted to generate self-explanations had higher gain scores (pre- to post-test), had deeper understanding, and answered a greater number of the harder questions correctly, than students who were not prompted to self-explain. The amount of gain was the same for both the high and low ability students (as measured by the California Achievement Test). Thus, the benefit of self-explaining can be elicited by using simple content-free prompts. Content-free prompts are powerful because they can be employed across different content domains, and they can be implemented easily as a general interface.

A similar result was obtained by King (1994). King taught peer-learning groups to ask questions using generic "question stems" that linked the to-be-learned material to their background knowledge. Although these question stems themselves were content-free, the peers often had to add content words in order to complete these question stems. King found that groups who used the question-stem prompts learned more than the control groups on most post-test measures.

Because prompting students to give explanations either to themselves or to their peers is helpful for learning, the question thus arises: Can automated prompting be just as effective or must a human prompter be involved (either an experimenter, as in Chi et al., 1994, or a peer, as in King, 1994)? Because computers do not yet have full natural language understanding, automating prompting means that the prompts must be administered in a pseudo-random fashion (i.e., without a pedagogical basis). In the Chi et al. (1994) study, the experimenter prompted the student to self-explain when he or she became silent. However, it is also possible that the experimenter prompted at times when the experimenter thought that the student was confused. If this was true, then knowing when to prompt the student required the experimenter to understand what the student had said. Therefore, to discriminate whether prompting alone (and thereby motivating the students to be constructive) was important or whether the timing of the prompting was also important, two different prompting techniques were employed in this second experiment, one by a human and one by a computer. The text materials, procedure, and interface used are identical to the first experiment. However, the read-only condition was not included because the critical comparison was between human and computer prompting.

Thus, the main difference between the first and second experiments is the introduction of content-free prompting. For the human prompting, a methodology from research on human-computer interaction was used. The prompter and student were put in a Wizard of Oz situation because they were seated in separate rooms, and the student was not told in advance whether he or she was speaking to a human or computer (Dahlbäck, Jönsson, & Ahrenberg, 1993). The reason for withholding this information was to control for the effects of talking with a human versus a computer, over and above the differences in prompting. The traditional use of the Wizard of Oz methodology is for "rapid prototyping" where a computer system is designed to interact as if it were human (Maulsby, Greenberg, & Mandler, 1993). A similar orientation was taken in the present study, but with a slight difference. Instead of prototyping a potential computer tutoring system, our intent was to empirically investigate specific features of a tutorial dialog that are hypothesized to be important for learning. In particular, it is conceivable that human prompters (such as tutors) are more likely to prompt when they think the students are confused or have incomplete understanding. Similar approaches have been taken by Fox (1991). Fox suggested that back-channel feedback, such as eye-gaze, facial expressions, and pauses are all linguistic devices used by human tutors to diagnose student confusion (for further support, see Anderson, 2002; Gluck, 1999).

The present learning situation restricted non-linguistic communication by asking a human to prompt a student through a dialog box. The prompter, however, was only able to ask the student questions that appeared at the top of a pre-configured list of prompts (for examples of the types of prompts used, see Appendix C of Chi et al., 2001). The effects of free-form typing on tutorial dialogs have already been shown to be effective (Hume, Michael, Rovick, & Evens, 1996), whereas constraining the tutor in a specific way has received little attention.

Design

Participants

Each condition included 10 students, for an overall total of 20 participants. All participants were selected from the same pool of undergraduate psychology majors, and they were given course credit upon completion of the experiment. None of the participants from the first study were recruited for the second experiment. Again, students who were selected to participate had adequate typing skills; however, no formal evaluation of typing speed was assessed.

Materials

Prompts. The prompts used in the second experiment were taken from a set of content-free prompts that were transcribed from actual tutorial dialogs (Chi et al., 2001). All content words were removed from the prompts. An example of a content-free prompt includes, "Could you explain how that works?"

Conditions

The computer interface used in the first experiment was modified to include a method for prompting the student to self-explain while learning. The student interface was basically identical to the one used in the first experiment, except that a field in the middle of the screen, labeled "Messages," periodically displayed a prompt selected by either a human or the computer, depending on the experimental condition. How and when the prompt is displayed in the human prompting condition will be described in the next section.

Human prompting. During the learning phase of the experiment, a TCP/IP connection was created between the student's interface and the prompter's window. The conversational flow was as follows (see Appendix A for two protocol excerpts). First, the computer automatically presented a sentence from the text at the top of the student's and prompter's screen. After the student read the sentence, she had two options available. She could either type a comment to the prompter, or she was instructed to type "ok" to signal that she had nothing to say. When the student clicked the "submit" button, the prompter was then able to read what the student wrote. If the prompter decided to ask a question, he clicked on the "send" button, which was located next to the corresponding prompt. The prompter's communicative abilities, however, were constrained because he was *not* able to type in a free-form way. Instead, he could only ask the student a question by sending a prompt from the top of a list of content-free prompts. The prompter was free to decide when he wished to prompt the student, although he was told to encourage students to self-explain if they did not type many comments for several consecutive sentences. When the prompter was finished, he clicked on a different button to tell the student to go onto the next sentence.

Automatic prompting. The automatic prompting condition used a computer program called the automatic prompting system (APS).[1] The dialog pattern found in the human prompting condition also holds for the automatic prompting group, and it was used to create the sequence of prompting for this condition. During the human prompting condition, a log of the dialog was generated for each individual sentence. The automatic prompting condition was yoked to the human prompting condition in terms of the number of prompts for each sentence on a sentence-by-sentence basis. For example, if the human prompter gave Student A two prompts for sentence 14, then the computer program (APS) also gave two prompts at sentence 14 to Student B.

The reason for yoking the students on a one-for-one basis across the two conditions was to determine if the prompts that were generated in the human prompting condition were tailored for the student at that particular time. We assumed that the dialog patterns for each individual should be different because each person has a different understanding of the circulatory system. For example, suppose Student A in the human prompting condition said she understood that lungs re-oxygenate the blood. She might not need a prompt during the passage that deals with oxygenation (e.g., sentence 15-16). However, if Student B, from the automatic prompting condition, believed the blood enters the lungs to supply oxygen *to* the lungs, then that student's dialog will be different. Because the background knowledge of each student is different, our assumption was the observed dialog patterns reflected those differences. Therefore, the prompting obtained in Student A's protocol should not help Student B. The differences in learning should reflect a difference between well-timed prompting, versus arbitrary (or "automatic") prompting.

Results

Equivalent Groups

Both human prompting and automatic prompting groups were equivalent in terms of their self-reported SAT-verbal scores ($t(18) = 1.22$; $p = 0.24$), pre-test vocabulary ($t(18) = 0.22$; $p = 0.83$), and pre-test mental model ($p = 1.00$, Wilcoxon Two-sided Test). The amount of time spent during the learning phase of the experiment was the only reliable difference between the human ($M = 49.7$ min.) and automatic prompting ($M = 32.5$ min.) conditions ($t(18) = 3.79$, $p = 0.001$). One reason why the groups were different could be attributable to an artifact of network traffic, as well as the reading and response time of the prompter. Because the computer stored the content-free prompts in memory, the time between the student input and computer response was nearly instantaneous.

Pre-test to Post-test Learning

The human ($M = 23.1$, $SD = 11.67$) and automatic ($M = 25.0$, $SD = 14.90$) prompting groups both increased their vocabularies from pre- to post-test (see Table 4). However, the groups did not differ from each other on the post-test vocabulary test.

Table 4
Pre-test and Post-test Performance on the Vocabulary Test for the Human Prompting (HP) and Automatic Prompting (AP) Conditions

		Pre-test		Post-test		
Condition	*n*	*M*	*SD*	*M*	*SD*	*p value for paired t-test*
HP	10	49.3%	12.3	72.4%	16.1	0.0002
AP	10	51.7%	16.9	75.7%	20.6	0.0005

Note. There were no reliable differences between the two groups for either the pre- or post-test.

Similarly, 70% of the human and the automatic prompting groups started the experiment with a Single-Loop model of the circulatory system. After the learning phase, the two groups did not differ on their post-test mental model drawings. The gain between pre-test and post-test mental models were equivalent in the two groups ($p = 0.92$, Wilcoxon Two-sided Test), essentially 80% of the students acquired the Double-Loop model (see Table 5).

Table 5
Percentage of Mental Models During Pre-test and Post-test for the Human Prompting (HP) and Automatic Prompting (AP)

Mental Model Category		Pre-test		Post-test	
Name	#	HP	AP	HP	AP
No Loop	1				
Ebb and Flow	2				
Single-Loop	3	70%	70%		10%
Single-Loop w/ Lungs	4			10%	
Multiple Loops	5	30%	20%	10%	10%
Double-Loop (1)	6		10%	60%	30%
Double-Loop (2)	7			20%	50%

Note. Definitions for the mental models can be found in Table 4 of Chi et al. (1989). Double-Loops (1) and (2) are the most scientifically accurate models of the circulatory system.

Similarly, both groups performed equally well on the Verbatim and Integration questions, and there was no interaction between condition and question type. The post-test measures for the prompting conditions were the same for Verbatim and Integration learning. Because the students in the automatic prompting condition learned as effectively as students who were prompted by humans, the two different types of prompting were collapsed into one group, called the prompted self-explanation (PSE) condition.

Correlation of Self-explanation and Learning

There was a reliable correlation between the amount of self-explanations and Integration learning ($r = 0.38$, $p = 0.10$, one-tailed test), but not Verbatim learning for the prompted self-explanation condition. On the other hand, paraphrasing was not correlated with either Integration ($r = -0.14$, $p = 0.56$) or Verbatim ($r = -0.10$, $p = 0.66$) learning. The results found here provide direct evidence for the claim, made in earlier studies (Chi et al., 1989; Chi et al., 1994), that paraphrasing is not a constructive activity and thereby does not contribute to overall learning of a conceptual domain. At best, it contributes to verbatim knowledge, as shown in Experiment 1.

Discussion

The evidence from the second experiment suggests that learning from a computer interface can be effective when prompted either from a human or computer. Furthermore, self-explaining significantly increased when it was prompted, rather than occurring spontaneously. For instance, when the data from Figure 1 (spontaneous self-explanation) is replotted against the amount of prompted self-explanation condition (see Figure 2), it is clear that the students generated more self-explanation when prompted. However, paraphrasing remained high and did not differ between experiments.

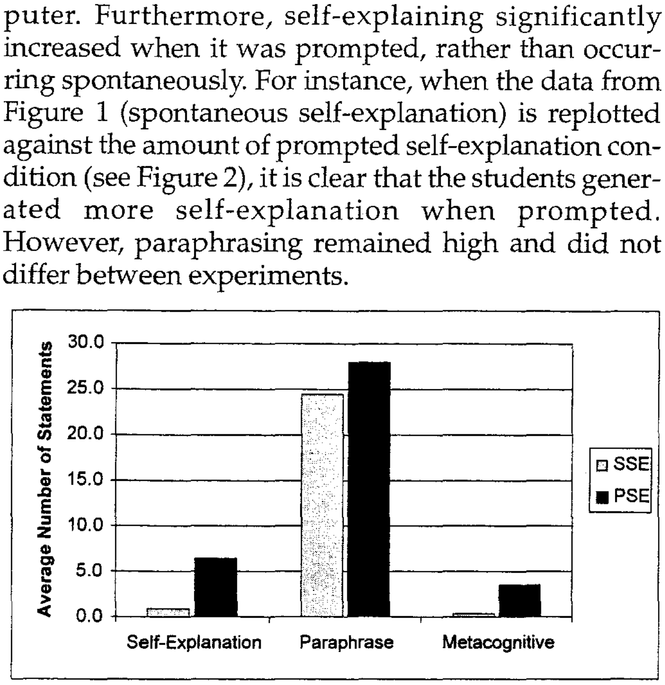

***Figure 2.* The average number of statements made by participants during the learning phase of the study in the spontaneous self-explanation (SSE) and prompted self-explanation (PSE) conditions.**

The effect of self-explaining with the computer interface was particularly effective for the students who produced a large amount of self-explanations. When restricting the analyses to only the upper quartile of self-explainers (n=5 participants) in the prompted self-explanation condition, they produced more self-explanations (M = 12.00, SD = 2.55) than the lower three quartiles of self-explainers (M = 4.67, SD = 2.38) ($t(18)$ = 5.87, $p < 0.0001$). The amount of self-explanations for the upper quartile correlated strongly with Integration (r = 0.92, p = 0.03) but not with Verbatim learning (r = 0.18, p = 0.77), thus replicating the effectiveness of self-explanations, especially for deep understanding (Chi et al., 1994). Although only 12.0 self-explanations were produced on average by the high explainers (which was still significantly less than prior studies on oral self-explanation), a significant correlation with deep learning was observed.

These results are encouraging for two related reasons. The effects of prompting did not seem to depend on whether the prompting came by way of a computer or human; therefore, implementing an automatic prompter is trivial and can be incorporated into many learning environments. Even though the prompting system does not have natural language understanding, and was administered in a somewhat arbitrary fashion, learning benefits were still observed. Although a formal analysis of the prompter was not conducted, he did provide a few justifications for prompting the students. The conditions under which the prompter asked a question included times when the student was relatively passive, when the student did not integrate across sentences, and when the prompter thought the student might be able to anticipate the next topic in the text. A talk aloud protocol from the prompter would be an interesting methodological extension of the present study.

General Conclusion

When given the opportunity and some general encouragement, students in earlier studies were found to spontaneously generate a great number of self-explanations orally. Table 6 summarizes the number of explanations generated orally in two different studies (Chi et al., 1989; Wathen, 1997), using basically the same text materials as used here (see Table 6). In contrast, when given the opportunity and some general instructions to type their explanations, the students in the first experiment, when left to their own devices to learn conceptual material, generated very few explanations. Instead, they predominantly paraphrased the text. According to prior research, it is possible that students paraphrase the to-be-learned material just as much as the students in the current study (for example Table 3 from Chi et al., 1989 shows that 32% of the successful and 42% of the less successful students' utterances were paraphrases). However, the clear difference is the relative lack of self-explaining here.

Table 6

A Comparison of the Amount of Spoken or Typed Self-Explanations Across Studies Using the Same Material

Study	*n*	*Frequency*	*Percentage*
Chi et al. (1994)			
Low	4	29.0	28.7%
High	4	87.0	86.1%
Wathen (1997)			
Talk-Individuals	18	34.2	69.8%
Present Study			
SSE	10	1.0	1.6%
PSE	20	6.5	10.3%
High PSE	5	12.0	19.4%

Note: The percentage of self-explanations generated equals the average frequency divided by the number of opportunities to self-explain (i.e., the number of text sentences). *SSE*=spontaneous self-explanation; *PSE*=the combined human and automatic prompting groups; and *High PSE*=upper quartile of the prompted self-explanation group.

The contrast between spoken self-explaining and typed self-explaining demonstrates that the amount of typed self-explanation is still far below what we might expect for spontaneously spoken self-explaining. This is most dramatically demonstrated by the fact that the highest quartile of self-explainers from Experiment 2 generated 1.5 times less self-explanation statements than the lowest group from Chi et al. (1994). The difference between spoken and typed self-explaining suggests two future studies. The first could compare average against expert typists, and the second could contrast verbal self-explaining against typed self-explaining.

The reason why students do not interact with the material in a deep way might be explained by the resource demands that typing places on the student. For example, expert typists are able to correct grammar, check the spelling, and even carry on conversations while transcribing a document, whereas the novice transcribers cannot (Gentner, 1988). In contrast, oral speaking is an automated task in which less time is spent on planning and execution. Therefore, self-explaining might be easier to do when speaking than during typing.

Besides the labor and resource intensive nature of keyboarding, there may be other reasons why students avoid self-explaining and prefer to paraphrase instead, as compared to actually speaking. One reason might be that keyboarding provides a record of what the students typed. Their input into the computer appears on the screen after they are finished typing, and it is inferred from the experimental instructions that their log files will be analyzed later. To avoid errors, the students may have preferred the safety of a paraphrased entry.

Although students paraphrased the material extensively, were we successful in motivating the participants to type self-explanations while using a computer interface? It seems that the answer depends upon the implementation used to elicit the self-explaining. Students may not type self-explanations spontaneously (e.g., the first experiment), as one might anticipate from the prior literature from spoken self-explaining. However, when prompted, students were able to increase their typed self-explanations, which then correlated with learning. This was most effective for those who did a lot of self-explanation (i.e., the high self-explainers).

The results from the second experiment compliment the findings from a study conducted by Aleven and Koedinger (2000), using a computer interface that also allowed free-form user input. They found that 36% of the students' statements were self-explanations. One reason why students in the study by Aleven and Koedinger produced more self-explanations than the present study might be attributed to the number of opportunities the participants received. Specifically, Aleven and Koedinger prompted their students an average of 34 times ($SD = 17$) per two sessions, while the present study prompted the students an average of 15.5 times. This further suggests that the number of self-explanations in a given learning session can increase when the number of opportunities is also increased (in the form of prompting). In other words, it does not matter *when* or *where* in the text one is prompted because one can always do some integration, repair, or inferencing any time, which yields an increase in learning. The determining factor may have been the absolute number of prompts given to the student. Therefore, one potential explanation for why there was a consistent gain in learning for both human and automatic prompting groups in Experiment 2 is because both groups received the same number of prompts (recall the yoking procedure).

By allowing students to type their self-explanations in a free-form way motivates students to be more constructive. One of the reasons why self-explanation is hypothesized to be effective is because the student generates new knowledge for him- or herself. Menu-drive systems may work for some, but not all. The results from the second study suggest that allowing students to type free-form explanations may be a useful way to motivate those who previously did not display learning gains from menu-drive systems.

References

Aleven, V., & Koedinger, K. (2000). The need for tutorial dialog to support self-explanation. In C.P. Rose & R. Freedman (Eds.), *Building Dialogue Systems for Tutorial Applications. Papers from the 2000 AAAI Fall Symposium* (pp. 65-73). Menlo Park, CA: AAAI Press.

Aleven, V., Koedinger, K., & Cross, K. (1999). Tutoring answer explanation fosters learning with understanding. In S.P. Lajoie & M. Vivet (Eds.), *AIED-99 Artificial Intelligence in Education: Vol. 50. Open learning environments: New computational technologies to support learning, exploration, and collaboration* (pp. 199-206). Amsterdam: IOS Press.

Anderson, J.R. (2002). Spanning seven orders of magnitude: A challenge for cognitive modeling. *Cognitive Science, 26,* 85-112.

Chi, M.T.H. (2000). Self-explaining expository texts: The dual processes of generating inferences and repairing mental models. In R. Glaser (Ed.), *Advances in instructional psychology* (pp. 161-238). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Chi, M.T.H., & Bassok, M. (1989). Learning from examples via self-explanations. In L.B. Resnick (Ed.), *Knowing, learning, and instruction: Essays in honor of Robert Glaser* (pp. 251-282). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Chi, M.T.H., Bassok, M., Lewis, M.W., 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., DeLeeuw, N., Chiu, M.-H., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science, 18,* 439-477.

Chi, M.T.H., Siler, S.A., Jeong, H., Yamauchi, T., & Hausmann, R.G. (2001). Learning from human tutoring. *Cognitive Science, 25*(4), 471-533.

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

Clark, H.H. (1999, September 1-3). *Speaking in time.* Proceedings of the ESCA workshop on dialogue and prosody, Veldhoven, the Netherlands.

Conati, C., & VanLehn, K. (1999). Teaching meta-cognitive skills: Implementation and evaluation of a tutoring system to guide self-explanation while learning from examples. In S.P. Lajoie & M. Vivet (Eds.), *AIED-99 Artificial Intelligence in Education: Vol. 50. Open learning environments: New computational technologies to support learning, exploration, and collaboration* (pp. 297-304). Amsterdam: IOS Press.

Conati, C., & VanLehn, K. (2000). *Further results from the evaluation of an intelligent computer tutor to coach self-explanation.* Paper presented at the ITS 2000, 5th International Conference on Intelligent Tutoring Systems, Montreal, Canada.

Dahlbäck, N., Jönsson, A., & Ahrenberg, L. (1993). *Wizard of Oz studies-Why and how.* Paper presented at the 1st ACM International Workshop on Intelligent User Interfaces.

Fox, B.A. (1991). Cognitive and interactional aspects of correction in tutoring. In P. Goodyear (Ed.), *Teaching Knowledge and Intelligent Tutoring* (pp. 149-172). Norwood, NJ: Ablex Publishing Corp.

Gentner, D.R. (1988). Expertise in typewriting. In M.T.H. Chi, R. Glaser, & M.J. Farr (Eds.), *The nature of expertise* (pp. 1-21). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Gluck, K.A. (1999). *Eye movements and algebra tutoring.* Unpublished doctoral dissertation, Carnegie Mellon University, Pittsburgh, PA.

Howie, D.E., & Vicente, K.J. (1998). Making the most of ecological interface design: The role of self-explanation. *International Journal of Human-Computer Studies, 49*(5), 651-674.

Hume, G., Michael, J., Rovick, A., & Evens, M. (1996). Hinting as a tactic in one-to-one tutoring. *The Journal of the Learning Sciences, 5*(1), 23-47.

Jeong, H. (1998). *Knowledge co-construction during collaborative learning.* Unpublished doctoral dissertation, University of Pittsburgh, Pittsburgh, PA.

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*(2), 338-368.

Lebie, L., Rhoades, J.A., & McGrath, J.E. (1995). Interaction process in computer-mediated and face-to-face groups. *Computer Supported Cooperative Work, 4,* 127-152.

Maulsby, D., Greenberg, S., & Mandler, R. (1993). *Prototyping an intelligent agent through Wizard of Oz.* In ACM SIGCHI Conference on Human Factors in Computing Systems, Amsterdam, The Netherlands.

Renkl, A. (1997). Learning from worked-out examples: A study on individual differences. *Cognitive Science, 21*(1), 1-29.

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

Turkle, S. (1984). *The second self: Computers and the human spirit.* New York: Simon and Schuster.

Wathen, S.H. (1997). *Collaborative versus individualistic learning and the role of explanation.* Unpublished doctoral dissertation, University of Pittsburgh, Pittsburgh, PA.

Weizenbaum, J. (1966). ELIZA - A computer program for the study of natural language communication between man and machine. *Communications of the ACM, 9*(1), 36-45.

Author Note:

Funding for this search is provided by the National Science Foundation, Grant Number NSF (LIS): 9720359, to the Center for Interdisciplinary Research on Constructive Learning Environments (CIRCLE, http://www.pitt.edu/~circle). The authors are indebted to the following for providing helpful comments on earlier drafts: Stephanie A. Siler, Randi A. Engle, Marguerite Roy, Rod D. Roscoe, Steve Jones, and Leslie R.M. Hausmann. We would also like to thank Anders Weinstein for programming the APS and Mark U. McGregor for being our wizard.

Footnotes:

[1] In describing the APS, it is also important to make clear what it is *not*. The computer program used in the current study does not employ any ability to parse or produce an utterance. A second feature that the APS does not have is the ability to craft a response from the user's input. A famous computer program, called ELIZA (which was initially developed to act as a Rogerian therapist), is an example of a non-intelligent computer program that creates a response based on the user's input (Turkle, 1984; Weizenbaum, 1966). Our automatic prompting system is even less sophisticated than ELIZA because the APS will not generate a prompt based on the user's input. The original ELIZA is able to crudely match the input from the user to keywords in her lexicon.

[2] The numbers on the left side of the dialog represents the line number taken from the protocol. To clarify the role of student and tutor, the student will be called "she" throughout the paper, while the tutor will be referred to as "he." The reason for the gender assignment is pseudo-random because the prompter used in our study was male.

Direct correspondence to:

Robert G.M. Hausmann, M.S.
Michelene T.H. Chi, Ph.D.
Center for Interdisciplinary Research on Collaborative Learning Environments
Learning Research and Development Center
University of Pittsburgh
Pittsburgh, PA 15260
email: bobhaus@pitt.edu or Chi@pitt.edu

Appendix A

An example of a student answering the prompter's content-free question.[2]

204 58. Renal circulation is the subsystem of systemic circulation that moves blood through the kidneys and back to the heart.
205 S: ok, but why to the kidneys?
206 T: Could you elaborate on what you just said?
207 S: I don't understand why there's a system of the systemic circulation that directly relates to the kidneys
208 T: Please click the "Next" button.

An example of a participant who did not answer the prompter's question.

87 30. Movement of gases and nutrients takes place across the thin capillary walls mostly from areas of greater concentration to areas of lesser concentration. This process is called diffusion.
88 S: diffusion - moving greater\ concentration to lesser concentration thru capillary walls
89 T: Is there anything else that you want to say about that?
90 S: no
91 T: Please click the "Next" button.