When Are Tutorial Dialogues More Effective Than Reading?

Kurt VanLehn[a], Arthur C. Graesser[b], G. Tanner Jackson[b],
Pamela Jordan[a], Andrew Olney[b], Carolyn P. Rosé[c]

[a]*Learning Research and Development Center, University of Pittsburgh*
[b]*Fedex Institute of Technology, University of Memphis*
[c]*Human Computer Interaction Institute, Carnegie-Mellon University*

Abstract

It is often assumed that engaging in a one-on-one dialogue with a tutor is more effective than listening to a lecture or reading a text. Although earlier experiments have not always supported this hypothesis, this may be due in part to allowing the tutors to cover different content than the noninteractive instruction. In 7 experiments, we tested the interaction hypothesis under the constraint that (a) all students covered the same content during instruction, (b) the task domain was qualitative physics, (c) the instruction was in natural language as opposed to mathematical or other formal languages, and (d) the instruction conformed with a widely observed pattern in human tutoring: Graesser, Person, and Magliano's 5-step frame. In the experiments, we compared 2 kinds of human tutoring (spoken and computer mediated) with 2 kinds of natural-language-based computer tutoring (Why2-Atlas and Why2-AutoTutor) and 3 control conditions that involved studying texts. The results depended on whether the students' preparation matched the content of the instruction. When novices (students who had not taken college physics) studied content that was written for intermediates (students who had taken college physics), then tutorial dialogue was reliably more beneficial than less interactive instruction, with large effect sizes. When novices studied material written for novices or intermediates studied material written for intermediates, then tutorial dialogue was not reliably more effective than the text-based control conditions.

Keywords: Tutorial dialogue; Learning; Physics education research; Qualitative physics; Conceptual physics; Natural language dialogue; Latent semantic analysis

1. Introduction

A tutor often faces the decision of whether to just tell the student an explanation or to try to elicit the explanation from the student via a series of questions. For instance, suppose the stu-

---

dent asks, "Why does the heart go 'lub, dub'?" The tutor could simply give a full explanation as a monologue: "The heart has four chambers. Two are the powerful ventricles and two are the weaker atria. …" On the other hand, the tutor could elicit an explanation via a series of questions:

> Tutor: "Well, how many chambers does the heart have?"
> Student: "What's a chamber?"
> Tutor: "In this case, it's a muscular bag that collects blood then squeezes it out."
> Student: "Like the ventricycle?"
> Tutor: "Yes, the *ventricle* is one kind of chamber. The other kind is called an *atrium.* Now how many ventricles are there?"

This example of a tutorial monologue and a tutorial dialogue illustrates several potential advantages of the dialogue over the monologue:

- Dialogue allows the tutor to detect and repair failed communications. In this example, the student did not know what *chamber* referred to, so the monologue version may not have been as understandable as the dialogue version.
- Dialogue allows detection and remediation of incorrect student knowledge. For instance, this student had an incorrect pronunciation for *ventricle.*
- Dialogue allows the tutor to assess the student's level of knowledge, adding content to "fill in" apparent gaps in the student's knowledge. For instance, because the student did not know the term *chamber,* the tutor added a little extra information about the collecting and squeezing functions of heart chambers. Although this information was absent from the monologue version, the squeezing function was presupposed by the terms *powerful* and *weaker* in the monologue, so this student many not have understood what those modifiers meant in this context.

There are many other potential advantages that are not illustrated here. For instance, a dialogue demands the student's attention to answer the tutor's questions, whereas the student's attention is free to wander with a monologue.

Studies of actual tutorial dialogue have unveiled both commonalities and variations, as we discussed later in this article. One important dimension of variation is the ratio of tutor-only explanations to tutor-and-student explanations. We use *interactivity* to refer to this dimension. If the tutor basically lectures, the instruction has a low degree of interactivity. If the tutor attempts to elicit most of an explanation from a student, then the instruction has a high degree of interactivity.

Several studies have found that higher interactivity correlates with larger learning gains. Wood and Middleton (1975) studied young children learning how to assemble a puzzle and found that learning gains were largest with tutors who adjusted their elicitations so that students had just enough information to move forward. On the other hand, Wood and Middleton (1975) commented that three tutors relied almost entirely on "simply showing the child how to put everything together and then asking him to do the same. This approach was disastrous" (p. 188). Chi, Siler, Jeong, Yamauchi, and Hausmann (2001) analyzed transcripts of college students learning about the heart while studying a text with a tutor. Chi et al. (2001) found that

measures of student participation in the dialogue correlated strongly with posttest scores. Several studies of typed conceptual physics tutoring have found that the average number of words per student utterance correlated with learning (Katz, Connelly, & Allbritton, 2003; Litman et al., in press; Rosé, Bhembe, Siler, Srivastava, & VanLehn, 2003). A study (Core, Moore, & Zinn, 2003) of typed tutoring of basic electricity and electronics found that learning gains correlated with the proportion of words produced by the student. All these studies have suggested that when tutorial dialogues are more interactive, in that the students participate more, then the learning gains are larger.

However, correlation is not causation, so it could be that some third factor, such as the motivation of the student or verbal fluency, causes the student to both learn more and to participate more in the tutorial dialogue. Thus, several experiments have manipulated the interactivity of tutorial dialogues to see whether interactivity causes, either directly or indirectly, larger learning gains. Before reviewing this literature, we make two methodological points.

First, it is widely believed that the content of the instruction can make a large difference in learning gains regardless of how it is taught. Because we are primarily interested in the impact of interactivity on learning, the ideal experiment should control for content. That is, both the tutoring and the low-interactive instruction (e.g., a lecture or even just reading a text) should cover the same information, one way or another. Several techniques can be employed to control for content. One is to run the tutoring condition first and videotape the tutoring sessions; the low-interaction instruction consists of watching those videotapes. Another technique is to use computer tutors instead of human tutors because their content can be well controlled. There are other techniques as well. Nonetheless, only a few of the studies we review have attempted to control rigorously for content.

Second, we state the hypothesis under test, which we call the *interaction hypothesis:* When one-on-one natural language tutoring, either by a human tutor or a computer tutor, is compared to a less interactive control condition that covers the same content, then the tutees will learn more than the nontutees. To make the literature review small enough to be manageable, we have limited the tutoring to natural language as opposed to the formal languages of mathematical expressions, menus, mouse gestures, forms, and so forth used by many computer tutors.

We review studies relevant to testing the interaction hypothesis in three groups. The first group of studies have produced results consistent with the interaction hypothesis; they have shown that tutoring was reliably more effective than a low-interaction control condition. The second group of studies have produced null results and thus do not support the hypothesis. The third group of studies have produced ambiguous results—it is unclear whether they support the interaction hypothesis or not.

1.1. Studies that have supported the interaction hypothesis

Wood, Wood, and Middleton (1978) had a human tutor implement four different strategies for teaching preschool children how to assemble a complicated block structure. One strategy implemented the following rule: "If the child succeeds, when next intervening offer less help. If the child fails, when next intervening take over more control" (Wood et al., 1978, p. 133). The other strategies were less interactive. For instance, the least interactive strategy had the tu-

tor just demonstrate the to-be-learned procedure. As predicted by the interaction hypothesis, the most interactive tutoring strategy produced the best performance on a posttest.

Swanson (1992) compared the highly interactive tutoring strategy of Wood et al. (1978) to simply lecturing. As in the Wood et al. study, the same tutor implemented both forms of instruction, but Swanson's students were college students learning how lenses work. As predicted by the interaction hypothesis, the more interactive tutoring produced more gains.

Several studies have compared reading a computer-literacy textbook to natural language computer tutoring that was designed specifically to emulate the tutorial dialogues found during human tutoring (Graesser et al., 2003; Person, Graesser, Bautista, Mathews, & Tutoring Research Group [TRG], 2001). As predicted by the interaction hypothesis, the tutored students learned more than the students who studied the textbook for an equivalent amount of time. Moreover, unlike the preceding studies, which have not attempted to control for content, these studies used a computer tutor designed to present the same information as the textbook. However, the tutor and textbook were written by different authors. As we discuss later, their content may not have been completely identical.

Lane and VanLehn (2005) compared two versions of a tutoring system that focused on teaching novice programmers how to design a program before writing the code for it. In the interactive condition, the computer tutor conducted a typed dialogue with students that elicited a design from them while providing hints and occasional directive help. In the noninteractive condition, students read a text with exactly the same content as the tutorial dialogue. Although some posttraining measures produced null results, the tutored students exhibited improved ability to compose designs, and their behavior suggested thinking at greater levels of abstraction than students in the reading group. Thus, this experiment partially supported the interaction hypothesis.

Evens and Michael (2006, Section 10.2) studied students learning how to predict changes in cardiophysiological variables caused by various medical events. Evens and Michael compared expert human tutoring to a control condition in which students read a textbook that was written by the tutors to have the same content as they normally covered. On a posttest that assessed student's accuracy at making predictions, the tutored students did significantly better than the control students as predicted by the interaction hypothesis.

These six studies all produced results consistent with the interaction hypothesis. That is, tutoring, either by a human or a computer, produced higher learning gains than a low-interaction control condition. Note that the control conditions' training involved only reading a textbook or listening to a lecture/demonstration. The control students did not solve problems or answer questions during training. They simply read or listened. These low-interaction conditions are quite different from the ones that have been used in the next group of studies, which have failed to support the interaction hypothesis.

1.2. Studies that do not support interaction hypothesis

Evens and Michael (2006, Section 18.4) compared two computer tutors that helped students learn how to predict changes in cardiophysiological variables. Both tutoring systems covered the same content. Both gave students practice at making predictions and gave them feedback on their predictions. One computer tutor remedied incorrect predictions by printing an expla-

nation of the correct derivation of the prediction. The other computer tutor remedied incorrect predictions by engaging students in a typed natural language dialogue. Although the interaction hypothesis predicts that the dialogue-based remediation should produce more learning gains than the text-based remediation, there were no significant differences between the tutoring systems' gains. Notice that in both conditions of this study, the students made predictions and received feedback during training. In the Evens and Michael (Section 10.2) study reviewed earlier, the control condition students were asked only to read a textbook during training and were not asked to make predictions.

Chi et al. (2001) took advantage of the propensity of untrained tutors to lecture and first had a group of tutors work with tutees naturally. These tutors were then trained to be more interactive, for example, by using content-free prompting as much as possible. Analyses of the dialogues showed that the tutors did most of the talking when untrained, and students did most of the talking after the tutors were trained. Contrary to the interaction hypothesis, the learning gains of tutees in the two groups did not differ. These results are the opposite of those found by Swanson (1992) and Wood et al. (1978) who also used a high-interaction condition with human tutors who were trained to be interactive. However, in the low-interaction conditions of the earlier studies, the tutors had been instructed to lecture and demonstrate, whereas in the low-interaction condition of the Chi et al. (2001) study, the tutors were instructed to tutor naturally. This could explain why the early studies' results supported the interaction hypothesis, but the Chi et al. (2001) study's results did not.

Rosé, Moore, VanLehn, and Allbritton (2001) compared Socratic and didactic strategies for tutoring students on basic electricity, both conducted by the same human tutor. Analysis of the transcripts indicated that the Socratic tutoring was indeed more interactive than the didactic tutoring. However, students in the didactic condition were not entirely passive. They still answered questions and participated in discussions, albeit not as extensively as in the Socratic condition. Contrary to the predictions of the interaction hypothesis, the learning gains of the Socratically tutored students were not reliably different from those of the didactically tutored students. However, there was a trend in the expected direction, and the number of students (20) was small, so a Type II error was possible.

Katz et al. (2003) compared interactive human tutoring to reading a text. In particular, Katz et al. had a computer present a question and the student type in a paragraph-long answer. Students in the reading condition would then study a paragraph-long version of the correct answer. In contrast, students in the human tutoring condition engaged in a computer-mediated (typed) dialogue with an expert human tutor. The tutorial dialogue showed little lecturing, so the tutoring qualified as interactive. Contrary to the predictions of the interaction hypothesis, the tutored students did not learn more than the reading students. Once again, however, the low-interaction instruction was not totally passive because students alternated between studying paragraph-long correct answers and writing their own paragraphs. Thus, they had the opportunity during training to apply the knowledge and strategies acquired through reading. Perhaps this motivated them to study the correct answers harder.

Reif and Scott (1999) compared human tutors to a computer tutor. Both taught students to solve physics problems using Heller and Reif's (1984) problem-solving strategy. The computer tutor did not attempt to converse with students in natural language but instead had students fill in blanks and click on menus to step through a solution to the problem. It gave imme-

diate feedback on incorrect entries, so it is not entirely clear whether it was more or less interactive than the human tutors. Suffice it to say that the interaction was different. Nonetheless, the learning gains were the same, thus lending no support to the interaction hypothesis.

Rosé et al. (2003) compared computer-based natural language tutoring to reading multi paragraph explanations that were written to have the same content as a maximally long tutorial dialogue. The instruction consisted of 10 short lessons, each comprised of a dialogue or a few paragraphs and concluding with the students writing a short summary of what they had just studied. Contrary to the interaction hypothesis, the tutored students learned no more than students who read the content instead of interacting with the computer tutor.

Craig, Sullins, Witherspoon, and Gholson (2006) compared students learning computer literacy by either working with a computer tutor or watching the computer tutor lecture. The lecture was constructed by having the computer tutor present as a spoken exposition all the instructional components that it could possibly discuss while tutoring students. With this careful control of content, the tutees had the same learning gains as the students who watched the lecture. However, Craig et al. (in press) also ran several other lecture conditions. These had deep questions inserted into them. The questions were either spoken in the same voice as the tutor (like a rhetorical question) or in a different voice as if they had come from a tutee. Either way, the addition of deep questions made the lectures more effective; the learning gains were reliably larger than those achieved by tutoring. However, this could be due to additional content contained in the deep questions or a separable effect of incremental gains by deep questions. At any rate, these results provide no support for the interaction hypothesis, which predicts that tutoring should be more effective than lecturing.

1.3. Studies that may or may not support the interaction hypothesis

Craig, Driscoll, and Gholson (2004) conducted several experiments in which students learned computer literacy by either working with a computer tutor or watching a video of the same tutor working with a human tutee. In particular, a video of every tutee in the tutoring condition was shown to one nontutees in the comparison condition. Thus, content was equated in the two conditions. Unfortunately, it is not clear whether the tutees learned more than the nontutees. Experiments 1 and 2 of Craig et al. (2004) reported that tutoring was more effective than observing a video of tutoring. However, Experiments 1 and 2 of Craig et al. (2006) reported that the tutees and the nontutees had the same gains.

Aleven, Ogden, Popescu, Torrey, and Koedinger (2004) compared two geometry tutoring systems. Both had students answer geometry questions such as "which two angles in this figure can be proved equal?" One tutor had students justify their answer by selecting postulates from a menu. The other had students type in a justification and gave them hints and advice until they entered an acceptable one. The latter tutoring system was much more interactive. According to the interaction hypothesis, it should have been more effective than the menu-based tutoring system. However, the results were mixed. The posttest assessed three skills.

- On the skill of giving correct answers, the gains of the two groups did not differ.
- On the skill of determining whether a figure gave enough information to answer a question, the gains of the two groups did not differ.

- On the skill of typing in justifications, the students of the more interactive tutor were more competent than students of the menu-selection tutor. The justification skill was assessed by giving 1/3 point for a justification that had just the name of the rule (e.g., "supplementary angles"), 2/3 point for an incomplete statement of the justification (e.g., "the two angles add up to 180 degrees"), and full credit for a complete statement of the justification (e.g., "the two angles add up to 180 degrees so the angles are supplementary").

The menu-based group's justification scores were lower for two reasons. First, they usually answered with just the name of a rule, which is what they did during training but were instructed not to do on posttest. Second, when they did attempt to state more than the name of a rule, they seldom managed to state it completely and correctly, perhaps because this was their first opportunity to practice this skill. Aleven et al. (2004) concluded "The explanation format affects communication skills more than it affects students' problem solving skills or understanding, as evidence by the fact that there was no reliable difference on problem-solving or transfer items" (p. 447). Whether this experiment supports the interaction hypothesis depends on whether this particular communication skill, as assessed in this particular manner, is considered an instructional objective of the geometry curriculum.

1.4. Discussion of the prior studies

To summarize, when the results have a clear interpretation with regard to the interaction hypothesis, they exhibited the following pattern:

- If students in the control condition engaged in no interaction at all and only read text or watched a lecture/demonstration, then interactive tutoring usually elicited larger learning gains than the control instruction as predicted by the interaction hypothesis (Evens & Michael, 2006; Graesser et al., 2003; Lane & VanLehn, 2005; Person, Graesser, Bautista, et al., 2001; Swanson, 1992; Wood et al., 1978). The only exception is the Craig et al. (2006) study, which compared working with a computer tutor to watching the tutor lecture.
- If students in the control condition both read text and tried to use the text's content to solve problems or answer questions during training, then interactive tutoring was usually not more effective than the comparison instruction (Chi et al., 2001; Evens & Michael, 2006; Katz et al., 2003; Reif & Scott, 1999; Rosé, Moore, et al., 2001; Rosé et al., 2003).

However, null results are often open to many interpretations, so the preceding conclusions merit further exploration. In particular, confusing patterns of null and positive results can be caused by aptitude–treatment interactions. High-competence students often learn equally well from many types of instructions, whereas low-competence students often learn better from more scaffolded instruction (Cronbach & Snow, 1977). When an aptitude–treatment interaction exists, experiments can have either null results or positive results depending on the prior competence of their students.

One reaction to the messy pattern of results is simply to decide that the interaction hypothesis is not worth more testing. However, the interaction hypothesis lies at the heart of many learning theories, public policies, and technological developments. For instance, many socially oriented theories of student learning emphasize the interaction between students and

teachers or students and more capable peers (e.g., Collins, Brown, & Newman, 1989; Vygotsky, 1978). Recent federal policy in the United States seems to view tutoring as one panacea for underperforming schools.[1] Major efforts in educational technology, such as natural language intelligent tutoring systems and computer supported collaborative work, assume that the interaction hypothesis holds far and wide. We acknowledge that the interaction hypothesis is much more specific than these broad issues because the hypothesis merely motivates a scientific prediction that contrasts natural language tutoring with lecturing or reading on the same content. However, solid results showing that tutoring ties with reading, even in a constrained setting, would begin to challenge some common assumptions.

Indeed, we began this research in 1999 following the assumption that the interaction hypothesis was essentially correct. Many of the negative results reviewed previously had not yet been discovered. Our main interest was to compare the effectiveness of two computer tutors, but we added in two control conditions—reading text and human tutoring. We assumed that human tutors would be more effective than reading, and that the computer tutors would fall somewhere in between. We were surprised to find out that all four conditions produced the same learning gains in our initial experiment. We eventually discovered that a more complex but systematic picture emerged after completing seven experiments. Because this article is an evaluation of the interaction hypothesis and not the computer tutors per se, we describe the computer tutors only briefly and refer to other publications that specify their tutoring mechanisms in more detail.

Our summary conclusion is that the interaction hypothesis holds in some situations but not in others. One factor that needs to be considered is whether the control condition is completely noninteractive. A completely noninteractive condition would include pure reading or video watching without any problem solving or question answering. As suggested by the results we reviewed earlier, tutoring seems to be more effective than such noninteractive conditions. A second factor, suggested by the results we describe later, is whether the students have adequate preparation to learn from the less interactive instruction. In particular, if novices read text that was written for intermediates, they learn less than tutees struggling through the same intermediate-level content with a human tutor to help them. However, there are conditions in which interactive tutoring is not effective. The value of interactive tutoring over reading text is minimal or nonexistent when (a) content is controlled, (b) students are required to answer questions as they study the text, and (c) the students are studying text written to their level (i.e., novices study text written for novices, and intermediates study text written for intermediates). Although our data support this conclusion, we conducted all our experiments in a specific framework covering Newtonian physics, which we describe in the next section. More research is needed from outside our framework and subject matter to test the generality of the results.

2. The experimental framework

Graesser, Person, and Magliano (1995) observed a pervasive dialogue pattern in human tutoring that had the following five steps:

1. The tutor poses a question or problem.
2. The student attempts to answer it.
3. The tutor provides brief evaluative feedback.

4. The tutor and student collaboratively improve the answer or solution. This can involve a moderately long dialogue.
5. The tutor ends the discussion often by asking the student if they understand and almost always getting a positive response.

An example of the five-step frame follows:

1. Tutor: What does a *t* test tell you?
2. Student: It tests whether a mean is significant.
3. Tutor: Sorta.
4. Tutor: Can it be applied to experiments with just one group, or do you need two or more groups?

   Student: More than one.
      Tutor: Right. Because the *t* test compares the means of the two groups. What does it tell you about the two means?[2]
   Student: Whether they are significant.
      Tutor: Almost. What you care about is whether one mean is really and truly higher than the other or whether the difference between them is just an accident of sampling. Does the *t* test tell you anything about that?
   Student: Yes.
            <etc.>
5. Tutor: So do you understand the *t* test now?
   Student: Yes.
      Tutor: Good. Let's go on.

If the student gives a completely adequate answer at Step 2, then the tutor gives positive feedback at Step 3 and does not do Steps 4 and 5. Thus, we refer to Steps 4 and 5 as the remedial part of the five-step frame. Normal classroom interaction has only the first 3 steps, a dialogue pattern that is often called initiate-respond-evaluate (IRE) in classroom discourse research (Leme, 1990; Mean, 1979; Sinclair & Coulthard, 1975). That is, the teacher initiates a question or problem, a student responds with an answer, and the teacher evaluates the answer. It is well documented that human tutors produce more learning than classroom teaching (Cohen, Kulik, & Kulik, 1982).

Graesser et al. (1995) hypothesized that the effectiveness of tutoring over classroom instruction lies in the tutorial dialogue of the remedial part of the five-step frame. In essence, it is the multi turn interactive nature of that dialogue that accelerates learning. In particular, if the interaction hypothesis is applied to the remedial part of the five-step frame, it predicts that learning would be hurt if Steps 4 and 5 were replaced with a short lecture that told the student the correct reasoning as illustrated following:

   Tutor: What does a *t* test tell you?
Student: It tests where a mean is significant.
   Tutor: Sorta.
   Tutor: The *t* test is useful in experiments where there are two groups, and you are interested in whether the mean of one group is really and truly higher than the other or

whether the difference is just an accident of sampling. The *t* test looks not only at the numerical difference between the means but also at how widely or narrowly distributed the two groups are. <etc.>

Our primary research goal was to test the interaction hypothesis in the context of the Graesser et al. (1995) five-step frame. Specifically, how is learning affected when the remedial part of the five-step frame is implemented by having the student either engage in tutorial dialogue or read a text given that the student covers the same inferences during the dialogue and the reading? We used either interactive tutorial dialogue or reading to implement the remedial part of the five-step frame. The interaction hypothesis predicts that the tutorial dialogue will cause larger learning gains than the text.

We implemented the tutorial dialogue with both human tutors and computer tutors. Human tutors were expected to be more skilled and adaptive dialogue participants. However, the content covered by the computer tutors' dialogue was more easily equated with the content included in the text. We used two computer tutors: Why2-AutoTutor (Graesser, Person, Halter, & TRG, 2001; Graesser, Wiener-Hastings, Wiener-Hastings, & Chris, 1999; Person, Graesser, Chris, Pomeroy, & TRG, 2001) and Why2-Atlas (Jordan, Makatchev, Pappuswamy, VanLehn, & Albacete, in press; Jordan & VanLehn, 2002; Makatchev, Hall, Jordan, Papuswamy, & VanLehn, 2005; Makatchev, Jordan, & VanLehn, 2004; Rosé, Roque, Bhembe, & VanLehn, 2002; VanLehn et al., 2002). For each question, the computer tutors were designed to cover the same set of points as the text in which "covering a point" means the computer tutor either recognizes the point in the student's initial answer to the question (Step 2 of the five-step frame), successfully elicits the point from the student during remediation (Step 4), or fails to elicit the point and instead explains it during remediation. Before discussing this method in detail, the basic idea of a "point" needs to be defined by first describing the task domain and then describing a cognitive task analysis of it.

2.1. The task domain

Our interest was in the impact of the interaction hypothesis on learning during five-step frame tutoring when content is controlled, and we wanted to use both human and computer tutors. Therefore, we selected a tutoring task in which natural language seems integral to the task and yet the task did not place impossible demands on the technology. We selected qualitative physics as the task domain and explanation questions as the task. We present a typical qualitative physics "why" question (Step 1 of the five-step frame):

A lightweight car and a massive truck have a head-on collision. On which vehicle is the impact force greater? Which vehicle undergoes the greater change in its motion? Explain why.

We present three typical student answers (Step 2 of the five-step frame) following verbatim. The first one is correct but incomplete (it omits mentioning Newton's third law). The other two are incorrect. Notice the similarity of the language and the style of reasoning in all three essays:

- "The equation, f = ma, determines how the motion of an object of a certain mass will change when a certain force is applied. The same force is applied to both vehicles. But, the mass of the car is less, therefore the change in acceleration must be greater."

- "Assuming the acceleration for both vehicles is the same. With Newton's 2nd law, F = MA, the massive truck will have a much greater impact force than that of the car. Because the acceleration the for both vehicles is the same the mass of the truck is much greater making the force of the impact the truck has much great. The car will undergo a greater change in its motion because the force is not as great and the mass of the car is much smaller than the trucks. The car will most likely change direction in its motion."
- "If we assume the vehicles have the same velocity we know that the vehicle with the larger mass will have the greater force, following Newton's second law f = m*a. The velocity can be brought into play because the higher the velocity the higher the initial acceleration. The velocity is reflective of the acceleration of the objects. Because the force will be greater in the vehicle with the larger mass, in this case the truck the impact force will be greater upon the car. If the two forces were equal they would cancel each other out because they are in opposite directions. However in this case the greater force belongs to the truck and the car will have the greater impact force. The car will also have the greatest change in motion. It should be moved backward when it impacts with the truck. The excess force in the opposite direction will translate into the car being pushed backward."

This task was chosen for several reasons. First, the qualitative physics task domain was tractable from a knowledge representation viewpoint because successful cognitive models of qualitative physics problem solving have been built (de Kleer, 1977; Ploetzner & VanLehn, 1997). Second, qualitative physics questions are known to elicit robust, persistent misconceptions from students such as "heavier objects exert more force" (Hestenes, Wells, & Swackhamer, 1992). Thus, the remedial part of the five-step frame must deal not only with errors of omission in the student's answer, such as failing to mention Newton's third law, but also important and deep errors of commission, namely, those caused by misconceptions. We hypothesized that tutorial dialogue should be especially good, compared to reading, at handling errors due to misconceptions.

2.2. The cognitive task analysis

A cognitive task analysis identifies the precise pieces of knowledge that students should learn. It allows experimenters to design all training conditions to teach the same pieces of knowledge and all tests to assess the same pieces of knowledge.

Our cognitive task analysis was based on the cognitive modeling of Ploetzner and VanLehn (1997). The analysis distinguishes two levels of generality. The higher level consists of mathematical versions of physics principles, such as Newton's second law: "$\mathbf{F} = m*\mathbf{a}$, where $\mathbf{F}$ is the net force vector acting on an object, m is its mass, and $\mathbf{a}$ is its acceleration vector." The lower level of generality consists of qualitative versions of these principles such as "If the net force on an object is in a certain direction, then so is the acceleration of the object." For each mathematical version of the principle, there can be many qualitative versions. The studies of Ploetzner and VanLehn (1997) suggest that these qualitative versions are the unit of transfer, so in the rest of this article, *principle* always refers to a qualitative version of a principle.

It is overly ambitious to determine the true nature of all physics misconceptions (a lifelong project indeed), but it was necessary to make some assumptions about them to design this

study. For example, we needed training and testing problems that elicited approximately the same set of misconceptions. Thus, we assumed that the students' misconceptions could be expressed at approximately the same level of generality, precision, and granularity as the principles. For instance, one common misconception is that objects that have no forces acting on them will gradually slow down. This probably stems from experience in the real world, where almost every motion is opposed by invisible frictional forces. However, for our purposes, we can view this misconception simply as a false belief that can be paraphrased as "If a moving object has no forces acting on it, then it will gradually slow down." It is at the same level of generality and granularity as a correct principle such as "If a moving object has no forces acting on it, then its velocity will be constant."

Our cognitive task analysis merely identifies principles and misconceptions. It does not address the controversies surrounding their interrelationships and the cognitive processes involved in "removing" a misconception (Chi, Slotta, & de Leeuw, 1994; di Sessa, 1993; Ranney & Thagard, 1988; Reiner, Slotta, Chi, & Resnick, 2000; Slotta, Chi, & Joram, 1995). If a misconception is evident on a student's pretest but not evident on the student's posttest, then we treat this as a positive feature of the instruction. However, when comparing two types of training, A and B, it is logically possible that A could be better than B at reducing misconception occurrences on a posttest, but B could be better than A at "really removing" misconceptions. Thus, we used multiple types of tests (including far transfer and retention tests in Experiment 3) and multiple methods of scoring them, some of which ignored misconceptions and some of which counted them.

2.3. The training problems

We developed the actual list of principles and misconceptions addressed by the experiments as we selected the training problems. Starting from an initial set of 53 qualitative problems culled from textbooks and other sources, we conducted an iterative process of analysis, modification, and selection based on (a) handwritten solutions to the problems from 120 students and four experts; (b) analyses of transcripts of a physicist (RS) tutoring students on the problems over a "chat" connection; (c) a solution to each problem as a two-column proof using the Ploetzner and VanLehn (1997) analysis as informal axioms (these proofs insured that none of the problems required subtle "tacit" knowledge); and (d) development of an ideal answer essay, about a paragraph long, for each problem by a collaboration of two physicists and a cognitive scientist.

In this fashion, we developed 10 problems for tutoring. For each question, we decided on the important propositions that an ideal answer essay would contain. These propositions were called the *expectations* for that problem. We also anticipated which misconceptions were likely to be manifested in the explanations for each problem. There were an average of 4.6 expectations and 4.8 misconceptions per problem. Table 1 shows one of our questions and its ideal answer essay, expectations, and misconception applications. To sketch the scope of the task domain, Appendix A lists all the training problems and essay test problems.

Our pilot participants and the participants for our first 3 experiments had all taken a college physics course. The training materials and tests were designed to be at an appropriate level of difficulty for these students. In later experiments, we used these same materials with students

Table 1
Expectations and misconception applications for the truck–car problem

Category	Respective Text
Question	If a lightweight car and a massive truck have a head-on collision, upon which vehicle is the impact force greater? Which vehicle undergoes the greater change in its motion? Defend your answer.
Ideal answer	The force of impact on each of the colliding bodies is due to interaction between them. The forces experienced by these bodies are thus an action/reaction pair. Thus, in terms of Newton's third law of motion, these forces will be equal in magnitude and opposite in direction. The magnitude of the acceleration produced by a force on different objects is inversely proportional to their masses. Hence, the magnitude of the car's acceleration due to the force of impact will be much larger than that of the more massive truck. A larger magnitude of acceleration implies a larger rate of change of velocity, which may be interpreted as greater change in motion. Therefore, the car undergoes greater change in its motion.
Expectations	• The magnitudes of the forces exerted by A and B on each other are equal. • If A exerts a force on B, then B exerts a force on A in the opposite direction. • The same force will produce a larger acceleration in a less massive object than a more massive object.
Misconceptions	• A lighter/smaller object exert no force on a heavier/larger object. • A lighter/smaller object exerts less force on other objects than a heavier/larger object. • The force acting on a body is dependent on the mass of the body. • Heavier objects accelerate faster for the same force than lighter objects. • Action and reaction forces do not have the same magnitude.

who had not taken college physics courses. It is likely that these materials were quite difficult for the latter population.

2.4. The four tutors

In our experiments, we contrasted various groupings of four types of remediation: human tutors, Why2-Atlas, Why2-AutoTutor, and text. They all implemented a four-phase pattern:

1. The student was asked a qualitative physics question.
2. The student entered an initial essay that answered the question and explained the answer.
3. Potential flaws in the student's answer and explanation were covered. This phase of the pattern varied across the four types of remediation.
4. The student was shown the ideal answer essay, was allowed time to study it, and then the next problem was presented to the student.

Phases 1 and 2 correspond to Steps 1 and 2 of the five-step frame. Phases 3 and 4 correspond to the remedial part of the five-step frame. In the case of the tutors, Phase 3 involved dialogue. In Phase 3 of the nontutoring condition, students read text and then edited their initial essay. Thus,

Phase 3 always involved some interaction, albeit not much in the case of the nontutoring condition.

2.4.1. Common components of the user interface

Fig. 1 shows the common components of the user interfaces that were used by students in all four conditions. The actual student interface in each condition had a few minor differences from the one shown in Fig. 1, but they all worked essentially the same way. The student was presented with a qualitative question in the upper right window. The student typed an explanation (called the essay) into the lower right window and clicked on the "Submit" button. The tutor and student then discussed the essay. When it was the student's or the tutor's turn, the participant either edited the essay or entered text in the lower left window, then clicked on the Submit button. Every turn was added to the bottom of the upper left, scrollable window. Neither participant could interrupt the other nor see utterances until they were submitted.

2.4.2. Human tutors

In one condition of the experiment, students worked with human tutors. The participants were in a different room from the tutor and knew they were communicating with a human tutor.
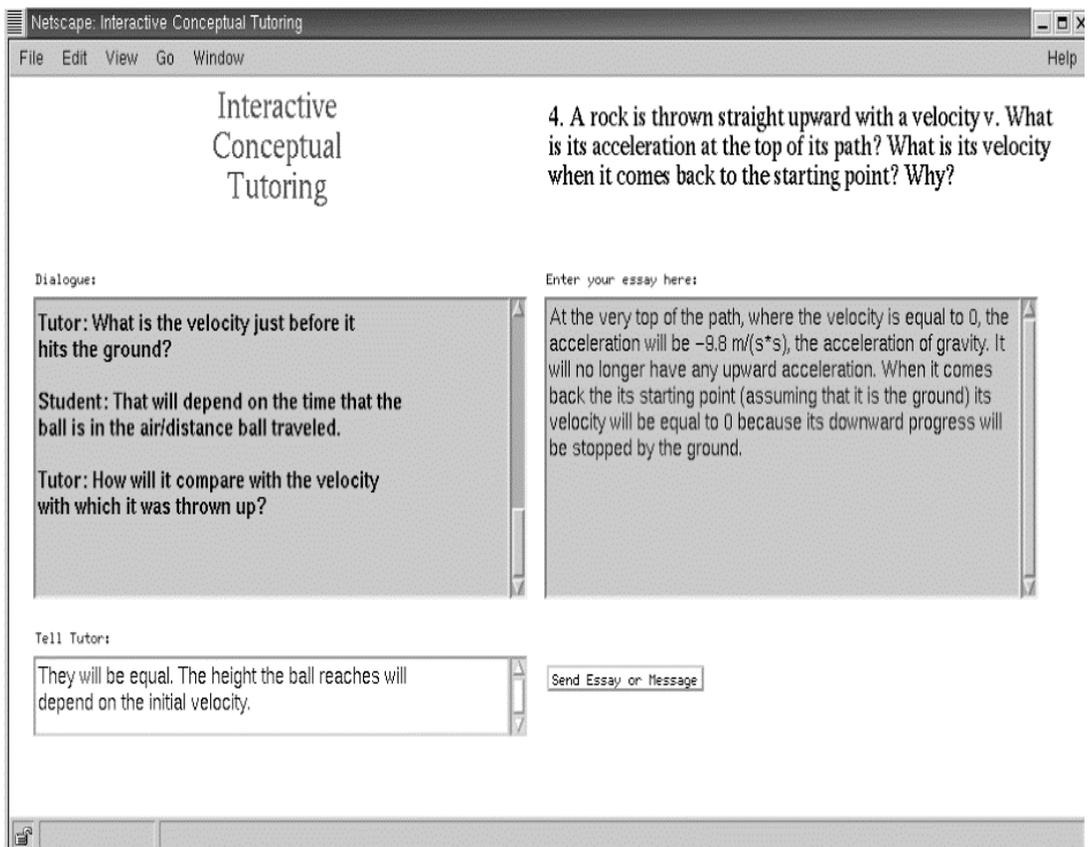

Fig. 1. User interface.

The students used the interface shown in Fig. 1. The human tutors used an identical interface that had just a few more buttons on it (e.g., for closing the current problem and opening the next one).

All tutors were instructed to elicit all the expectations for each problem and to be vigilant for a set of specific misconception applications for each problem. They were instructed to avoid lecturing the student. They knew that transcripts of their tutoring would be analyzed.

Many of the students in the human tutoring condition (e.g., 9 out of 21 in Experiment 1) were tutored by RS, a retired university physics professor working full-time with the project. The other three human tutors were university physics professors/instructors. All the tutors had spent many hours helping students individually. RS in particular had logged approximately 170 hr tutoring pilot participants using the computer-mediated setup.

2.4.3. Why2-Atlas

After a student in the Why2-Atlas condition entered an essay, the system analyzed it using a combination of knowledge-based and statistical natural language processing and reasoning techniques (Jordan, Makatchev, & VanLehn, 2004; Makatchev et al., 2005; Rosé et al., 2002). As we discussed earlier, every problem had a specific list of expectations and misconceptions for it that were based on the cognitive task analysis. If an essay had flaws, the tutor picked one (either a missing expectation or a misconception that seemed to be present) and discussed it with the student.

The discussion was organized as a knowledge construction dialogue (KCD). A KCD was based on a main line of reasoning that the tutor tried to elicit from the student by a series of questions. Typically, the tutor then summarized the main line of reasoning. This design was strongly influenced by Circsim-Tutor's directed lines of reasoning (Evens & Michael, 2006; Michael, Rovick, Glass, Zhou, & Evens, 2003). Table 2 shows an example of a KCD, with the

Table 2
A Knowledge-Construction Dialogue (KCD)

---

1. Tutor: Although it is clear that the truck exerts a force on the car, does the car also exert a force on the truck?
2. Student: I don't think so.
   a. Tutor: Could the car possibly dent the truck?
   b. Student: Not likely.
      i. Tutor: The car may not always dent the truck, but if the car was heavy enough and hit a soft part of the truck, then it could dent the truck.
   c. Tutor: If the car can dent the truck, then it must be exerting a force on the truck, and sometimes the force is enough to dent the truck.
3. Tutor: So when the truck and car collide, they each exert a force on the other. Whenever that happens, namely, two objects exerting forces on each other, the pair of force is called what?
4. Student: Action–reaction.
5. Tutor: The forces are an action–reaction pair. What is the relationship between their magnitudes?
6. Student: Same magnitudes but opposite directions.
7. Tutor: What is the relationship between the directions of the force on the car and the force on the truck?
8. Student: Opposite.
9. Tutor: So the car exerts a force on the truck, and at the same time the truck exerts an equal and opposite reaction force on the car.

---

lines numbered and indented to facilitate discussion here. In this case, the main line of reasoning has four steps, each of which is elicited by a tutor question (Lines 1, 3, 5, and 7) and summarized in the last tutor statement (Line 9). If the tutor failed to detect a correct answer to its question as in Line 2, the tutor initiated a subdialogue (Lines a through c). Different wrong answers could elicit different subdialogues. Subdialogues could be nested (e.g., Line i).

The dialogue management approach can be loosely categorized as a finite state model. However, it used a reactive planner that allows states to be skipped if the goal of a state was already achieved; it also backtracked and retried if a dialogue management plan failed (Freedman, Rosé, Ringenberg, & VanLehn, 2000). Tutor responses were specified in a hand-authored, push-down network (Jordan, Rosé, & VanLehn, 2001). State nodes in the network indicated that the system should either question the student or push or pop to other networks. The links exiting a state node corresponded to anticipated student responses to the question.

Anticipated student responses were recognized by looking for particular phrases and their semantic equivalents (Rosé, Jordan, et al., 2001). In the case of Table 2, Line 4, phrases that would be accepted as a correct answer were "third-law pair," "action–reaction pair," or "equal and opposite forces." However, it would also accept semantic equivalents such as "third law" or "a pair of third-law forces." KCD questions were phrased to invite short answers, but students sometimes provided long ones anyway. When the students provided short answers, the accuracy was high (Rosé, Jordan, et al., 2001). When students provided long answers, most of the words in the student's response were ignored. This could lead to disfluencies. For instance, the tutor detected "same" in line 6 and ignored the rest of the student's words, which makes its next question (Line 7) seem odd. Students sometimes appeared to detect this because they occasionally stopped elaborating their answers.

When a KCD for a flaw finished, the tutor asked the student to revise the essay. The student edited the text in the essay window and submitted it. If the student's modification fixed the flaw, then the cycle started again: The tutor searched for flaws in the newly revised essay, picked one, and discussed it. On the other hand, if the student's modification did not fix the flaw, the tutor said, "I'm sorry, either I wasn't clear enough or I didn't understand you. So that we can go on, I'll show you what I think summarizes the point. You need something like this in your essay. <text> When you are ready to continue, hit the Submit button."

This procedure could not be implemented on 5 of the 10 training problems due to lack of development time. In particular, the mechanisms for analyzing the student essays were not available on those 5 problems although the KCDs were. Thus, after the student entered an initial essay for one of these partially implemented problems, Why2-Atlas indicated that it would present some advice that was independent of the student's essay. It then ran through KCDs for each of the problem's expectations. After each KCD, it asked the student to revise the essay by saying some variant of "Although this won't necessarily fit smoothly into your essay, here's what I would have said at this point" and displayed a sentence or two for the essay. It did not check that the student entered this modification. Instead, it just went on to the next KCD when the student pressed the Submit button. Once all of the problem's KCDs had been presented, Why2-Atlas presented the ideal answer essay. When the student was done studying it, the tutor moved on to the next problem. Thus, for these 5 training problems, students worked through more KCDs than they probably needed. Although this defect affected the interpretation of the

Experiment 1 results, it was not an issue for the other experiments in which we used Why2-Atlas (5 and 7). On the later experiments, Why2-Atlas was able to analyze students' essays for all the training problems and to present only the KCDs that were relevant to defects in the student's essay.

2.4.4. Why2-AutoTutor

The Why2-AutoTutor students used a slightly different user interface, which is shown in Fig. 2. The window in the lower right quadrant accommodated whatever the student typed in during the current turn. The student's turn was processed with a speech act classifier (Olney et al., 2003) and a statistical natural language processing (NLP) technique called latent semantic analysis (LSA; Foltz, Gilliam, & Kendall, 2000; Graesser et al., 2000; Landauer, Foltz, & Laham, 1998). In the upper left window, an animated conversational agent had a text-to-speech engine that spoke the tutor's turns. It used facial expressions with emotions when providing feedback on the student's contributions. The agent also used occasional hand gestures to encourage the student to type in information. The recent dialogue history was displayed in a scrollable window in the lower left of the display. The tutor's most recent turn was added immediately after the tutor had finished speaking it.

Fig. 2. User interface of Why2-AutoTutor.

The interaction with Why2-AutoTutor was slightly different than the interaction between students and the other tutors. Although the other tutors asked the student to revise their essays after each flaw was discussed (and included a separate essay entry window), Why2-AutoTutor did not. Instead, it tried to elicit the correct expectations from the student. A particular expectation was considered missed if the match between it and the student's contributions did not meet a threshold according to an LSA cosine measure (Graesser et al., 2000; Olde, Franceschetti, Karnavat, Graesser, & TRG, 2002). If the contributions were missing an expectation, the tutor tried to elicit it from the student with prompts and hints (Graesser et al., 2001). If these attempted elicitations failed, then the tutor simply asserted the missing expectation. If the emerging explanation showed signs of an error or misconception, the tutor asked a question to verify the presence of the misconception and then attempted to correct it.

In addition to eliciting expectations and remedying misconceptions, Why2-AutoTutor had dialog moves that were designed to fulfill the goal of having a smooth, conversational, mixed-initiative dialog while still keeping the student focused on generating explanations for the physics problem (Person, Graesser, & TRG, 2002). For example, the tutor answered a student question by first classifying it into one of 20 different categories, including those in the Graesser and Person (1994) question taxonomy, and then accessing and displaying a paragraph of information from Hewitt (1987), a popular textbook on conceptual physics. However, students rarely asked questions as is the case in most learning environments (Graesser & Olde, 2003; Otero & Graesser, 2001), so the question answering facility probably had little impact. This mixed-initiative dialog was managed by an augmented state transition network called a dialog advancer network (Graesser et al., 2001; Person, Graesser, Kreuz, et al., 2001).

When Why2-AutoTutor assessed that all flaws in the student's beliefs about the problem had been remedied, three dialogue moves always occurred.

1. Why2-AutoTutor randomly selected one of the anticipated misconceptions for the problem and asked a diagnostic question. If the student answered correctly, the tutor moved on; otherwise, the tutor corrected the misconception in a single turn and went on. The purpose of the diagnostic questioning was to facilitate remedying misconceptions just in case the LSA analysis was unable to detect misconceptions.

2. To encourage mixed-initiative dialogue, Why2-AutoTutor invited the student to ask a question. After the student posed a question, the tutor answered it as described previously and went on.

3. The tutor asked the student to enter a complete essay. This essay was not used to control the multi turn dialogue. It served as the analogue to the essay that is eventually constructed in the other tutoring conditions.

Regardless of what is entered in the final essay, the tutor ended by presenting the ideal answer essay. When the student had studied it, the tutor moved on to the next physics problem. Presentation of the ideal essay was the last dialogue move in all conditions of the experiments.

It should be noted that many features of Why2-AutoTutor have direct correspondences in Why2-Atlas. Both tutoring systems implement a four-phase pattern:

1. The tutor presented a problem.
2. The student entered an initial essay within their first dialog turn.

3. The tutor identified flaws and attempted to correct all of them.
4. The tutor presented the ideal essay.

The dialog management of the third phase was somewhat different in the different tutors, but the content of the expectations and misconceptions were exactly the same. Whereas Why2-Atlas launched a KCD for a missing expectation, Why2-AutoTutor presented a series of hints and prompts until the student articulated the expectation. Both tutors simply asserted the expectation when they assessed that their elicitation techniques failed. When misconceptions were expressed by the student, Why2-AutoTutor directly corrected the misconception in a single turn, whereas Why2-Atlas launched a remedial KCD. However, Why2-AutoTutor, unlike Why2-Atlas, ended Phase 3 with the three dialogue moves listed previously.

2.4.5. Canned text remediation

The fourth condition of the experiment implemented a minimal-interaction form of instruction. It is called the canned text remediation condition because the software had the student enter an essay, read some canned text intended to remedy potential flaws in such essays, and then edit their essay. The canned text remediation condition did not analyze the student's essay, and it did not adapt the canned text to the student's essays flaws. All students saw exactly the same canned text.

More specifically, after the student entered the initial version of the essay, the canned text remediation presented "minilessons" to the student, one after another. Each minilesson consisted of one or more paragraphs of text. The minilessons were developed by converting the main line of reasoning of a Why2-Atlas KCD from dialogue to monologue. For instance, if the KCD asked "What direction is the gravitational force acting on the pumpkin?," and the correct answer was "Downward," then the minilesson had "The direction of the gravitational force acting on the pumpkin is downwards." In converting the KCD dialogues to minilesson monologues, we tried to keep the content as similar as possible.

Appendix B shows the minilessons for the truck–car problem. As it illustrates, the minilessons repeated the same point several times with different wording because the point is relevant to different expectations and different misconceptions. Moreover, if several problems addressed the same misconception, the corresponding minilessons were presented for each one. Thus, the canned text remediation was more redundant than a typical textbook and specifically addressed the problem the student had just attempted to answer.

2.4.6. Summary

The following is a list of the main phases of all conditions:

1. The student read a qualitative physics question.
2. The student typed in an essay that answered the question.
3. The students engaged in an activity intended to get them to learn physics while removing flaws in their essay in which a flaw is either a misconception or other error that is present or an expectation that is absent.
    a. In the human tutoring condition, the student participated in a typed dialogue with a human tutor. The student occasionally edited the essay, often because the tutor asked them to.

    b. In the Why2-Atlas condition, the tutor selected a flaw in the student's essay, conducted a KCD with the student about that flaw, and then asked the student to edit the essay to fix the flaw. This repeated until the essay had no flaws.

    c. In the Why2-AutoTutor condition, the tutor selected a flaw in the student's essay and conducted a dialogue to remedy it. This continued until all flaws have been discussed. The student then answered a question about a randomly selected misconception, was invited to ask Why2-AutoTutor a question (which it answered), and entered a final version of the essay.

    d. In the canned text remediation condition, the student read a minilesson for every possible flaw regardless of whether it occurred in the student's essay or not. Then the student was asked to edit the essay to remove any flaws it might have.

4. The student was shown a paragraph-long ideal answer to the problem.

3. Experiment 1

3.1. Design

We randomly assigned participants to one of four training conditions: human tutor, Why2-Atlas, Why2-AutoTutor, and canned text remediation. We assigned the students in the human tutoring conditions to one of four different tutors. All participants worked on the same problems in the same order, but the amount of time they took completing them varied according to their abilities. The main dependent variables were the pretest scores, the posttest scores, and the time to complete the 10 training problems. We administered a posttraining attitudinal survey, but we do not report the results here.

3.2. Participants

The participants were 98 university students who were currently taking or had recently taken introductory college physics but had not taken advanced physics courses or mechanical engineering courses. If the students were currently taking college physics, then they must have taken their first midterm because it covered the main topics of the tutoring (kinematics and forces). The participants were volunteers responding to advertisements at the University of Pittsburgh, the University of Memphis, Christian Brothers College, and Rhodes College. Students were compensated with money or extra course credit.

Although 98 students showed up for their first sessions, 6 dropped out, and another 6 were deleted because the data were unusable after equipment failure. The initial assignment of students to the various conditions was human tutor ($N = 21$), Why2-Atlas ($N = 26$), Why2-AutoTutor ($N = 26$), and canned text remediation ($N = 25$). The corresponding numbers of students after attrition were 18, 22, 24, and 22, respectively.

3.3. Materials

We described the 10 training problems and their development earlier, and they are presented in Appendix A. Two physics tests, A and B, were also developed. Half the students

received Test A as a pretest and half received Test B as a pretest. The other test was used as a posttest. Each version of the test (A and B) consisted of 4 essay problems and 40 multiple choice problems. We wrote the essay problems on the tests (see Appendix A) to address the same principles and misconceptions as the training problems and to require no other knowledge of physics.

For each expectation covered in the training problems, we wrote one multiple-choice question. For each misconception that could appear during training, we wrote one or more multiple-choice questions, and many were adaptations of ones that appear on the Force Concepts Inventory (Hestenes et al., 1992), a standard test of physics misconceptions. Appendix C shows several multiple choice problems from Test A. This sample of test problems all address Newton's third law, which is one of the main points of the training problem described in Table 1. Like the Force Concepts Inventory, our test problems probe the same concept in many different situations to gauge the generality of the student's knowledge and to elicit situation-specific misconceptions.

3.4. Procedure

All students filled out a consent form, filled out a background questionnaire on their physics courses, took a pretest, went through one of the four training conditions, took a posttest, and completed an attitudinal questionnaire. We presented the training problems in the same order for all students. We administered the tests and the training on computers in laboratories. The experimenters were either present in the laboratories or were nearby to facilitate initial use of the system and to restart the software if it crashed. All the software was Web based so that participants in Memphis could use software running in Pittsburgh and vice versa. We limited the sessions to at most 4 hr to prevent fatigue. Most students completed the study in two sessions. They typically completed 5 training problems in the first session and the other 5 training problems in the second.

3.5. Did all four tutors teach the same content?

Before discussing the results of the experiment, it is important to assess whether the different training conditions covered approximately the same content. Because we designed the three computer-based training conditions, it was comparatively easy to insure content-equivalence for them. However, we had no such control over the human tutors. If the human tutors covered different material than the other tutors, for example, then it would not be surprising if human tutors ended up having different scores on learning outcome measures. Therefore, it is important to verify that the human tutors actually did cover the content that we intended them to cover.

It would be impractical to analyze the content of tutoring sessions on several levels of discourse content, structure, and cohesion, but we did perform some analyses that provided some assessment of content equivalence. We used LSA to compare the content of the tutors' contributions during training. Just as Why2-AutoTutor uses LSA for its conceptual pattern-matching algorithm when evaluating whether student input matches the expectations and misconceptions, it is possible to use LSA to assess the similarity of tutors' content among the various training sessions. LSA is a widely used, high-dimensional, statistical technique that among

other things measures the conceptual similarity of any two pieces of text such as a word, sentence, paragraph, or lengthier document (Foltz et al., 2000; Graesser et al., 2000; Landauer et al., 1998). LSA converts a bag of words to a point in a multidimensional space. The cosine distance between two points can be interpreted as a measure of the content similarity of the two texts that the points represent (Graesser et al., 2000; Landauer et al., 1998). Although it is possible for these LSA cosine similarity scores to be slightly negative, reasonable values vary from 0 (no overlap) to 1 (perfect similarity).

For each student, we collected the bag of words presented by the student's tutor and converted that bag to a point in the LSA space. One can visualize this as three clusters of points: one for the human tutees, one for the Why2-AutoTutor tutees, and one for the Why2-Atlas tutees. The size of the cluster corresponds to the within-condition similarity of the text read by the students. The mean cosine scores are shown along the diagonal of Table 3. As one would expect, the computer tutees had compact clusters (high similarity of texts: .927 and .940), and the human tutees had a more dispersed cluster (lower similarity: .711). The canned text was the same for all students in that condition, so its cluster was a set of identical points.

The off-diagonal entries of Table 3 indicate content similarity across conditions. Each off-diagonal entry reports the cosine distance between a point in one condition and a point in another condition averaged over all such pairs of points. One can visualize these off-diagonal values as the distance between clusters.

To determine whether the human tutors were conveying content that was similar to the content of the other conditions, we needed to interpret the cosine distances in terms of a few intuitive "benchmark" similarities. The following is one possible set of benchmarks:

- Some dialogue dissimilarity is introduced by the students' contributions, which drive the dialogue in different directions. Why2-AutoTutor and Why2-Atlas had within-condition cosine distances of .927 and .939, which suggests that a value of .930 is a very high value for dialogue similarity and indicates how much dissimilarity is due to the students.
- We created the canned text by converting the maximally long dialogues of Why2-Atlas into text, so this is intuitively the next highest degree of similarity. The corresponding cosine is .845.
- We designed the two tutoring systems to cover the same content, but the templates that drive the tutor turns were written by different authors, and the system's dialogue manage-

Table 3
LSA cosine similarity score means (*M*) and standard errors (*SE*) from Experiment 1

Tutor	Human		Why2-Atlas		Why2-AutoTutor		Canned Text Remediation	
	M	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*
Human	.728	.019	.685	.020	.707	.021	.659	.018
Why2-Atlas			.927	.014	.686	.021	.845	.021
Why2-AutoTutor					.939	.009	.677	.020
Canned text remediation							1.00	.000

Note. LSA = latent semantic analysis.

ment was different. Thus, their between-condition similarity, .686, represents the next lowest degree of similarity. The similarity of the canned text to Why2-AutoTutor, .677, is close, as one would expect.

Against this background, the top row of Table 3 (i.e., Ms = .685, .707, and .659 in columns 4, 6, and 8, respectively) compares the human tutorial dialogues content to the other three conditions. The average, .683, is approximate the same as the third of our benchmarks previously. That is, the similarity of the two computer tutors with each other is about the same as the similarity of the human tutors with the computer tutors. These LSA analyses support the claim that the content and inferences in the human tutoring condition were similar to the content covered in the other conditions. We point out that this LSA-based measure and its intuitive scaling was used again in a later experiment to assess content similarity.

3.6. Multiple-choice test scores

The interaction hypothesis implies that the human and computer dialogue-based tutors should elicit more learning than the canned text remediation. We used both multiple-choice and essay tests to measure learning gains. In the next few sections, we present these results.

We analyzed the multiple choice tests by scoring the test items as right or wrong and then converting to proportion correct. Table 4 shows mean pretest scores, posttest scores, and the adjusted posttest scores for each condition along with the standard errors. Adjusted posttest scores refer to posttest scores that have had the pretest score factored out in an analysis of covariance (ANCOVA). The pretest scores were not reliably different among the four conditions, $F(3, 82) = 1.23$, $p = .31$, mean standard error (MSE) = .031, so the students in the various conditions started out on an even playing field with respect to incoming competence. In the analysis of variance (ANOVA) with the Condition × Test Phase factorial design, there was no significant main effect of experimental condition and no significant Condition × Test Phase interaction. However, there was a robust main effect of test phase, with mean posttest scores being significantly higher than mean pretest scores, .772 versus .645, respectively; $F(1, 82) = 99.78$, $p < .05$, $MSE = .013$. Therefore, there were robust learning gains in the four conditions. In an ANCOVA with multiple-choice pretest scores as the covariate, the adjusted posttest scores of the conditions were not reliably different overall, $F(3, 82) = 0.59$, $p = .62$, nor were there significant pairwise differences. In summary, the students in all four groups learned and learned about the same amount. These results do not support the interaction hypothesis.

Table 4
Multiple-choice test means (M) and standard errors (SE) in Experiment 1

Dependent Measure	Human		Why2-Atlas		Why2-AutoTutor		Canned Text Remediation	
	M	SE	M	SE	M	SE	M	SE
Pretest	.596	.042	.702	.038	.650	.036	.640	.038
Posttest	.736	.031	.812	.028	.759	.027	.785	.028
Adjusted posttest	.767	.021	.782	.019	.759	.018	.791	.019

It is conceivable that the students in all four conditions found the material so easy or so hard to learn that their posttest scores were at ceiling or at floor. Either of these possibilities would explain the null effect of the experimental conditions. The pretest and posttest scores in Table 4 measure the proportion of correct responses. These values are clearly not close to 1.0 nor 0.0. However, it may be that some of the items were impossible to answer, whereas others were nearly always answered correctly. When we excluded posttest items that were always answered incorrectly or always answered correctly, then the highest posttest mean was .82, which is clearly not at ceiling or floor. Even when we excluded items that were answered correctly by less than 25% of the students and correctly by 90% or more of the students, the highest posttest mean was .81. We reran the ANCOVAs with data that excluded the very difficult items and the very easy items in an attempt to concentrate on the items that were most malleable to learning gains. However, the learning gains across conditions were still not reliably different. In short, it is difficult to account for the lack of significant differences among conditions by appealing to ceiling or floor effects.

As Cronbach and Snow (1977) noted, instruction can be much more effective for low-competence students than for high-competence students who are often able to learn equally well from all kinds of instruction. To check for such an aptitude–treatment interaction, we divided the students into low- and high-prior-competence groups using a median split on their pretest scores. A 2 (low vs. high pretest) × 4 (four conditions) ANOVA showed no reliable Condition × Competence interaction, which suggest that an aptitude–treatment interaction was not present. However, it is important to acknowledge that the students had completed the relevant parts of a physics course. This may have restricted the range of competence and limited our ability to detect aptitude–treatment interactions.

3.7. Near versus far transfer

It is possible that the lack of differences between conditions is due to their encouraging different kinds of learning rather than different amounts. Perhaps all four kinds of tutoring elicited near transfer, but only the interactive tutors elicited far transfer. We did not intentionally design the test items to vary along the near versus far dimension, so we divided the multiple-choice problems post hoc according to their similarity to the training problems. We ran a 2 (near vs. far) × 4 (four conditions) ANCOVA on the posttest scores with pretest scores as the covariate.

As one would expect, students tended to have higher adjusted posttest scores with the near-transfer items than the far-transfer items, $F(1, 175) = 18.20$, $p < .001$ for the main effect. For the far-transfer problems alone, the adjusted posttest means across conditions were not reliably different, $F(3, 84) = 1.06$, $p = .37$. For near-transfer problems, the conditions' adjusted posttest means were not significantly different, $F(3, 84) = 0.41$, $p = .75$. In summary, all four conditions showed approximately the same gains for both near- and far-transfer multiple choice problems, contrary to the predictions of the interaction hypothesis.

3.8. Coding the essays for expectations and misconceptions

We coded the essay tests to determine which expectations and misconceptions were present in each essay and whether they were explicitly mentioned in the essay or only implied by the

essay. We used multiple coders and an elaborate training procedure, with moderate intercoder reliability. From the coding, we extracted three scores:

- Stringent expectation score = the proportion of expectations coded as explicitly present.
- Lenient expectation score = the proportion of expectations coded as explicitly or implicitly present.
- Lenient misconception score = the proportion of misconceptions coded as explicitly or implicitly present, a lenient criterion. Low scores are better.

Table 5 shows the results of all these scoring techniques. Each mean is followed by the standard error.

The pretest scores did not differ significantly among the four experimental conditions when we performed a one-way ANOVA on each of the three dependent measures. In the ANOVA with a factorial design between tutoring condition and test phase, there was a main affect for test phase, suggesting that all participants learned between pretests and posttests, but there were no significant effects of condition and no significant interactions between condition and test phase. In an ANCOVA with pretest scores as the covariate, none of the adjusted posttest scores were reliably different from the others, suggesting that all the groups learned the same amount, contrary to the predictions of the interaction hypothesis.

It was worth checking to see if the null effect could be explained by ceiling or floor effects. The expectation scores are the proportion of expectations possible for a problem that the student articulated in the essay. Because the scores were not close to 1.0 or 0.0, ceiling and floor effects were unlikely. The misconception scores are the proportion of anticipated misconceptions that appeared in the essays. These numbers were low on the posttest and not much lower than the pretest. A floor effect is conceivable, but it would not explain the significant difference between pretest and posttest.

Table 5
Essay test score means (*M*) and standard errors (*SE*) for Experiment 1

Scoring	Human		Why2-Atlas		Why2-AutoTutor		Canned Text Remediation	
	M	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*
Stringent expectation								
Pretest	.120	.028	.155	.025	.133	.024	.128	.028
Posttest	.271	.049	.331	.044	.275	.042	.380	.044
Adjusted posttest	.283		.315		.276		.385	
Lenient expectation								
Pretest	.296	.053	.452	.057	.295	.046	.330	.048
Posttest	.523	.064	.546	.057	.540	.055	.627	.055
Adjusted posttest	.546		.494		.564		.634	
Lenient misconception								
Pretest	.063	.012	.041	.010	.061	.010	.071	.010
Posttest	.032	.010	.033	.009	.030	.008	.021	.009
Adjusted posttest	.032		.033		.030		.021	

3.9. Holistic scoring of the essays

Our essay coders sometimes expressed the view that they could tell how good the student's answer was even though they could not easily code it for expectations and misconceptions. Thus, two physics instructors jointly defined a grading rubric and then graded all the test essays using a standard letter-grade scale (A–F). We call this the *holistic* scoring of the essays. The interjudge reliability was high ($\alpha = .89$). The means and standard errors appear in Table 6. The letter grades were linearly transformed; so 1.0 was the maximum grade, and 0.0 was the minimum. As with the other learning measures, an ANOVA showed a reliable main effect for test phase but no main effect for condition and no interaction. In an ANCOVA with pretest scores as the covariate, the adjusted posttest scores were not significantly different across conditions. Thus, this measure also shows that students in all four conditions improved significantly from pretest to posttest, but they improved by approximately the same amount, contrary to the predictions of the interaction hypothesis.

3.10. Combined scores

The tests were designed so that the same principles were used on many multiple-choice and essay problems. The same misconceptions could also arise in many places. If we use only the conventional score, we risk overcounting principles that can occur in several places and undercounting those that can only occur in one place. It may be more informative to count each principle and misconception once regardless of how many opportunities it had to appear during testing. Moreover, this analysis would combine evidence from the multiple-choice test with evidence from the essay test. This combined analysis was feasible because the multiple-choice and essay tests were already analyzed in terms of expectations and misconception applications. Although we omit the details here, this analysis also showed that students in all conditions learned (their posttest scores were significantly higher than their pretest scores in all conditions), but they learned approximately the same amount (an ANCOVA on posttest scores, with pretest scores as the covariate, showed no differences across conditions). These results do not support the interaction hypothesis.

3.11. Efficiency

The canned text remediation presented minilessons for every possible flaw in the student's essay, whereas the tutors remedied only the flaws that they found in the student's essays. Thus,

Table 6
Holistic essay score means (*M*) and standard errors (*SE*) in Experiment 1

Dependent Measure	Human		Why2-Atlas		Why2-AutoTutor		Canned Text Remediation	
	M	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*
Pretest	.322	.065	.492	.057	.355	.045	.362	.060
Posttest	.600	.077	.603	.068	.593	.055	.678	.065
Adjusted posttest	.638	.063	.540	.058	.610	.053	.690	.055

Table 7
Means (*M*) and standard errors (*SE*) of work and wait times (minutes) in Experiment 1

Time Spent	Human		Why2-Atlas		Why2-AutoTutor		Canned Text Remediation	
	M	SE	M	SE	M	SE	M	SE
Working	120	10.2	160	9.2	183	8.8	61	9.2
Waiting	154	5.5	7	5.0	5	4.7	0	5.0
Total	275	12.5	167	11.3	188	10.8	61	11.3

it is possible that the advantage of interactivity lies more in its efficiency: how much time is required to bring the student to mastery. In this section, we discuss training times among the four conditions.

Table 7 shows the total time that students spent on the training problems as well as segregating that time into time they spent actually working (e.g., the time between the end of the tutor's turn and the end of the student's turn) and the time spent waiting for the tutor (i.e., the time between the end of the student's turn and the end of the tutor's turn). The times are in minutes with standard errors.

A one-way ANOVA on work time was statistically significant, $F(3, 82) = 34.18$, $p < .05$, $MSE = 18.70$. Post hoc pairwise comparisons were consistent with the following pattern of scores: canned text remediation < human tutoring < Why2-Atlas = Why2-AutoTutor. The canned text remediation students spent about an hour because they were simply reading the minilessons and entering two essays per problem, whereas students in the other sections spend about 2 or 3 hr because they had to do much more typing to interact with the tutors.

The students in the human tutoring condition spent more than half their total elapsed time waiting for the human tutor to type in responses. The wait time per tutor turn was 24.4 sec for the human tutoring condition versus 1.5 sec for Why2-Atlas and 1.6 sec for Why2-AutoTutor. These long wait times may have reduced the effectiveness of the human tutoring condition.

3.12. Discussion of Experiment 1

Multiple measures of learning gains all suggested exactly the same conclusion: Students in all four conditions gained significantly, but they all gained about the same amount. The measures were (a) scores on multiple-choice tests, (b) holistic scoring of essay tests (assigning A through F letter grades), (c) componential scoring of the essay tests (coding for expectations and misconception applications), and (d) combined scoring of multiple-choice and essay tests (coding for principles and misconceptions). We checked for ceiling effects and floor effects, aptitude–treatment interaction, and for differential transfer, but once again, no differences between the training conditions emerged. Although it is always difficult to interpret null results, we believe that in this case, students in all four conditions actually did learn about the same amount contrary to the predictions of the interaction hypothesis. Moreover, when we conducted statistical power analyses, there was sufficient power in the likelihood of finding a significant effect given there was in fact an effect (.90 or higher, assuming an effect size between two conditions of 1 sigma).

It was quite unexpected that our computer tutors did so well compared with human tutors. The computer tutors often did not recognize correct assertions and misconceptions expressed by the students due to the limitations of language understanding technology. Yet the computer tutors still performed quite well, on par with the humans. Although we are gratified that our computer tutors were just as effective as human tutors, before we are ready to celebrate, we need to understand why the canned text remediation students learned just as much as the students of all three kinds of tutors.

When the time students spend waiting for the tutor's response was removed, the canned text remediation students worked about 1 hr, the human tutored students worked about 2 hr, and the computer tutored students worked about 3 hr. The rapid progress of the canned text remediation students makes sense because they spent most of their time reading and only had to enter two essay-length answers per problem, whereas the other students had to type in answers to the tutors' questions as well as entering essays. The computer tutors probably required more time from their students than the human tutors because the computer tutors often failed to recognize correct assertions made by the student and thus required the students to restate them.

If the null results of Experiment 1 indicate a true tie in learning gains, then the fact that the canned text students spent much less time in training than the tutored students suggests that interaction merely slows students down without helping them learn more. This interpretation contradicts conventional wisdom in the learning sciences. Clearly, more experimentation is needed. In Experiment 2, we used a different kind of control condition; in Experiment 3, we used different assessments of learning gains; in Experiments 4, 5, 6, and 7, we used students with less prior knowledge of physics. In Experiment 4, we used the same materials as used in Experiment 1; in Experiment 5, we used somewhat simpler materials; and in Experiments 6 and 7, we used considerably simpler materials.

4. Experiment 2: More conventional control conditions

The tie between Why2-AutoTutor and canned text remediation is all the more surprising when AutoTutor's earlier successes are considered. In the knowledge domain of computer literacy (Graesser, Lu, et al., 2004; Graesser et al., 2003; Person, Graesser, Bautista, et al., 2001), students of AutoTutor repeatedly learned more than students who studied a textbook for similar amounts of time. However, we enumerate following several major differences between Experiment 1 and these earlier studies of AutoTutor that could potentially explain the difference in their findings.

First, students in the canned text remediation conditions not only read text; they also answered 10 essay questions and then corrected their answers. In the earlier AutoTutor studies, students in the textbook conditions have merely read text during the training phase. They have answered questions only during pretesting and posttesting. The canned text remediation's alternation of reading with more active processes of essay writing and correcting may have increased students' engagement when reading as suggested by our review of the literature earlier. In particular, the canned text remediation students may have engaged in more self-explanation than did the computer literacy textbook students.

Second, Experiment 1 used a different task domain than the earlier AutoTutor studies. Perhaps the characteristics of the knowledge are quite different for computer literacy and physics. Qualitative physics is notorious for its persistent misconceptions, as we discussed earlier, whereas computer literacy is perhaps more amenable to conceptual change and less saturated with persistent misconceptions. If none of the students learned much qualitative physics and all held on to their naïve physics beliefs, this would possibly explain why all conditions produced the same gain. However, this explanation is not consistent with our finding that students in all conditions had significant learning gains, both in terms of increases in expectations and removal of misconceptions.

Third, Experiment 1 may have controlled for content more successfully than the earlier AutoTutor studies. The qualitative physics domain permits thorough cognitive task analysis and modeling (Ploetzner & VanLehn, 1997). This allowed us to author tutoring and canned text that addressed exactly the same principles and misconceptions. In contrast to qualitative physics, knowledge about computer literacy is more open-ended, incomplete, fragmentary, and unsystematic. This may have made it more difficult for the computer literacy authors to insure that AutoTutor's curriculum scripts covered the same content as the textbook. In fact, differences in learning gains between AutoTutor and textbook controls were smaller when information was removed from the textbook that was not directly relevant to AutoTutor's curriculum scripts (Graesser et al., 2003). If it were possible to perform the systematic cognitive task analyses that were conducted with qualitative physics and to use them to insure that curriculum scripts, textbook, and tests all addressed exactly the same computer literacy content, then perhaps AutoTutor would tie with the computer literacy textbook just as it tied with the canned text remediation in Experiment 1.

Last, the canned text was repetitious. The KCDs were written to be independent of their context. Thus, if a KCD attempted to cover an inference in the middle of a line of reasoning, it first had to review some of the reasoning leading up to that point because it could not assume that the preceding context had covered it. When all the KCDs for a line of reasoning were converted to minilessons, the later ones tend to repeat points made in the earlier ones. Moreover, if different problems addressed the same misconceptions, then the minilessons for the misconceptions were repeated. Thus, the canned text was more repetitious than the textbooks that had been used in the earlier AutoTutor studies. If repetition is important for learning, this too would explain why the canned text remediation students learned more than the textbook students. On the other hand, the canned text remediation's minilessons were simply concatenated, which may have hurt their global coherence.

We designed Experiment 2 to test some of these explanations for the difference between Experiment 1's results and the earlier AutoTutor results. We compared Why2-AutoTutor to a textbook condition in which students only studied the textbook without writing and revising essays. Although the textbook was designed to be equivalent to Experiment 1's canned text, their content, coherence, and repetitiousness may have differed, so we compared them with software that evaluates content, coherence, and dozens of other linguistic characteristics. If the key differences between the earlier experiments and Experiment 1 was that the canned text remediation students wrote essays or that the texts varied in content, coherence, and redundancy, then we would expect to see Why2-AutoTutor > textbook. On the other hand, if task domains of qualitative physics and computer literacy were fundamentally different, then we might find Why2-AutoTutor = textbook in qualitative physics.

4.1. Method, materials, and participants

The instructional objectives, training problems, and tests were the same as in Experiment 1 except that we did not give the attitude assessment. Why2-AutoTutor was the same except that it incorporated an improved method for deciding whether a student's turn included an expectation (Hu et al., 2003). We developed the textbook by selecting passages from Hewitt's (1987) *Conceptual Physics* textbook that covered the target instructional principles. We included a third condition wherein students received no instruction but merely took the pretest during the first session of the experiment and took the posttest during the second session.

The participants were drawn from the same population as in Experiment 1 except that the University of Pittsburgh volunteers were replaced with students from the University of Mississippi. The remaining students came from the University of Memphis and Rhodes College as in Experiment 1. We assigned students to conditions randomly but unevenly: Why2-AutoTutor ($N = 32$), textbook ($N = 16$), and no instruction ($N = 19$). There was a rationale for assigning approximately twice as many students to the tutoring condition as each of the other two conditions. We intended on conducting correlational analyses between student abilities and learning gains; such analyses require approximately 30 participants for a satisfactory statistical analysis. In this study, we did not focus on such correlational analyses, however.

4.2. Results

As in Experiment 1, the pretests and posttests contained a multiple-choice test and an essay test. We computed the proportion of multiple-choice questions that were answered correctly on the pretest and posttest. Table 8 presents the means and standard errors of the pretest and posttest scores in the three conditions.

We conducted an ANOVA on the scores using a 3×2 factorial design, with condition as a between-subject variable and test phase (pretest vs. posttest) as a repeated measures variable. There was a statistically significant Condition $\times$ Test Phase interaction, $F(2, 64) = 12.28$, $p < .01$, $MSE = .005$. The pattern of means clearly showed more learning gains from pretest to posttest in the Why2-AutoTutor condition than the other two conditions. An ANCOVA was statistically significant when we analyzed the posttest scores using the pretest scores as a covariate, $F(2, 63) = 14.81$, $p < .01$. The adjusted posttest scores showed the following ordering among means: Why2-AutoTutor > textbook = no instruction. The adjusted posttest score

Table 8
Multiple-choice test means (*M*) and standard errors (*SE*) for Experiment 2

Dependent Measure	Why2-AutoTutor		Textbook		No Instruction	
	M	*SE*	*M*	*SE*	*M*	*SE*
Pretest	.597	.029	.566	.040	.633	.037
Posttest	.725	.026	.586	.036	.632	.033
Adjusted posttest	.727	.016	.610	.022	.608	.020

for Why2-AutoTutor (0.727 ± .016) was similar to its score for Experiment 1 (0.759 ± .018), suggesting that the minor population difference had no effect.

The effect size of the learning gains of Why2-AutoTutor was 1.02 when the adjusted textbook mean served as the control. Although this effect size was greater than the earlier studies of AutoTutor, the pattern of results was similar. Person et al. (2001) reported AutoTutor > textbook = no instruction, with an effect size of .50 on an aggregated measure and an effect size of .28 on the deep questions measure that most closely approximated the measures used here. Graesser et al. (2003) reported AutoTutor > textbook > no instruction, with an effect size of .23 on the deep questions measure when the textbook condition was used as a control.

To examine a potential aptitude–treatment interaction, we split students into high- and low-prior-competence groups via a median split on their pretest scores. A 3 × 2 ANOVA showed no significant Condition × Competence interaction for the multiple-choice posttest scores. As in Experiment 1, there appeared to be no aptitude–treatment interaction in this experiment.

As in Experiment 1, we scored the pretest and posttest essays both holistically and by coding for expectations and misconceptions present in various degrees in the essays. Table 9 shows the means and standard errors. Although the adjusted posttest scores tended to be higher for the AutoTutor group than the other two groups, none of the between-condition comparisons were statistically reliable. The essay tests appear to have been a less sensitive measure of competence than the multiple choice tests.

The training times differed among conditions. The students participated in two sessions, answering half the training questions in each session. The AutoTutor students worked with the

Table 9
Essay test score means (*M*) and standard errors (*SE*) in Experiment 2

Dependent Measure	Why2-AutoTutor		Textbook		Nothing	
	M	*SE*	*M*	*SE*	*M*	*SE*
Pretest						
Stringent	.158	.02	.151	.03	.121	.03
Lenient expectation	.410	.05	.478	.07	.414	.07
Stringent misconception	.023	.007	.028	.010	.017	.009
Lenient misconception	.065	.013	.057	.018	.035	.017
Holistic	.303	.045	.290	.060	.308	.058
Posttest						
Stringent expectation	.250	.03	.166	.04	.106	.04
Lenient expectation	.573	.05	.484	.07	.404	.06
Stringent misconception	.025	.007	.034	.009	.024	.009
Lenient misconception	.049	.012	.083	.017	.076	.016
Holistic	.475	.045	.345	.050	.310	.060
Adjusted posttest						
Stringent expectation	.249	.03	.165	.04	.110	.04
Lenient expectation	.579	.04	.467	.06	.408	.06
Stringent misconception	.003	.01	.003	.01	.003	.01
Lenient misconception	.005	.01	.008	.02	.008	.02
Holistic	.465	.045	.348	.050	.308	.060

tutor for 63.0 min ($SD = 17.1$) during the first session and 63.0 min ($SD = 26.1$) during the second session. The textbook students read the text for 46.3 min ($SD = 24.1$) during the first session and 34.2 min ($SD = 17.9$) during the second session. A 2×2 ANOVA showed a main effect for condition $F(1, 45) = 18.02$, $p < .001$, $MSE = 561.5$. This suggests that the time on task was somewhat higher for the AutoTutor students than the textbook students.

4.3. The content, coherence, and redundancy of the instructional texts

We intended that the textbook would cover the same content as Why2-AutoTutor and the canned text remediation condition in Experiment 1 despite the fact that the textbook focused on introducing basic concepts and principles of mechanics, whereas the canned text and Why2-AutoTutor focused on applying them to solve qualitative physics problems. To put it in the terminology of quantitative task domains (e.g., Renkl & Atkinson, 2003)

- Why2-AutoTutor coached students through the solving of 10 problems and provided instruction on concepts and principles whenever the student seemed to need it.
- The canned text presented 10 worked examples of problem solving plus instruction on a concept and principle whenever it was used.
- The textbook introduced concepts and principles using simple worked examples for illustration.

Thus, all three sources contained a mixture of introductions to concepts and principles and applications of them during problem solving. By design, they should cover the same material in different ways.

To test potential differences in content among conditions, we performed the same LSA-based analysis as in Experiment 1. When we compared the content of the textbook condition with the contents of the four conditions of Experiment 1, the mean LSA cosine similarity score was .582. In particular, the cosine similarity for the textbook versus the canned text was .656. This was slightly below the third benchmark we discussed earlier (.686), which measures the content similarity of the two computer tutors. This suggests that the textbook and the canned text had only slightly different content.

As we discussed earlier, it was important to examine whether the textbook and canned text remediation content differed in coherence, redundancy, and readability. To evaluate this, we analyzed them with a computer tool called Coh-Metrix (Graesser, McNamara, Louwerse, & Cai, 2004). All eight Coh-Matrix coherence measures (argument overlap, stem overlap, local coreference, global coreference, causal cohesion, local LSA coherence, global LSA coherence, and connective frequency) showed an advantage for the canned text remediation content, which is statistically significant according to a Wilcox sign test ($p < .05$). Redundancy (measured by type-token ratios) was not statistically different. Standard readability metrics (Flesch Reading Ease Score, Flesh–Kincaid Grade Level and logarithm of content word frequency) also showed no differences between the canned text and textbook.

Thus, the canned text and the textbook appeared to differ in coherence and perhaps content but to have had similar redundancy and readability according to some standard measures. The difference is easily appreciated by reading the canned text in Appendix B. Each paragraph seems to make just a few points, but it makes them in several ways. This may explain why the

canned text was so effective in Experiment 1 as a control condition and why the textbook was inferior to AutoTutor in Experiment 2 and in previous studies.

4.4. Discussion of Experiment 2

The Why2-AutoTutor students had learning gains that exceeded those of students reading the textbook. Moreover, reading the textbook did not produce any learning gains at all, nor did the comparison condition in which students received no training. This latter result would be surprising for students and teachers who routinely rely on the reading of textbooks for knowledge acquisition in courses. However, it is not at all surprising for researchers who emphasize the importance of active construction of knowledge, problem solving, and explanations for the achievement of deep knowledge. Moreover, in both this experiment and the computer literacy experiments (Graesser et al., 2003; Person, Graesser, Bautista, et al., 2001), students had already read similar textbooks and had been tested on them. In summary, all of these results replicate the experiments conducted with AutoTutor in the domain of computer literacy.

What has yet to be determined, however, is why the canned text remediation condition did so well in Experiment 1 (equivalent to the computer and human tutors), whereas the textbook condition did so poorly in Experiment 2 (equivalent to no instruction at all and much worse than Why2-AutoTutor). There are at least three possible explanations for the discrepancy.

One explanation, which is consistent with the literature review we presented earlier, is that students in the canned text remediation condition alternated between writing essays and reading, whereas students in the textbook condition only read. Perhaps the activity of generating and correcting essays facilitated learning as would be expected by constructivist theories of learning. It may also have increased the amount of self-explanation while reading the canned text compared to reading the textbook.

A second explanation is that the textbook was less coherent than the canned text as indicated by eight measures of coherence in the Coh-Matrix software(Graesser, McNamara, et al., 2004). In several studies, learning gains have been affected by text coherence as defined by these measures (McNamara, 2001; McNamara, Kintsch, Songer, & Kintsch, 1996). However, the direction of the effect depended on the prior knowledge of the students. For low-knowledge students, low-coherence text produced smaller learning gains. For high-knowledge students, low-coherence text often produced larger learning gains. We found that the low-coherence textbook produced smaller learning gains than the high-coherence canned text, which is consistent with the results of McNamara et al. only if we assume our students had low prior knowledge. This is unlikely for three reasons. First, our students had taken physics before, had read similar textbooks before, and had taken tests on the relevant knowledge as part of their coursework. Second, the canned text was designed during extensive pilot testing to be at the appropriate level of difficulty for these students. Third, we did not find an aptitude–treatment interaction. Despite these difficulties, the fact remains that the textbook was measurably less coherent than the canned text, which might explain the difference in learning.

A third explanation involves content differences. The canned text and Why2-AutoTutor focused more on application of knowledge (i.e., problem solving) than introduction of it, whereas the textbook focused more on introduction than application. Thus, the canned text remediation, Why2-AutoTutor, and the tests all had similar content, whereas the textbook con-

tent was broader than the content of the tutors and tests, so broad that the participants may have diluted their study with auxiliary material not related to the tests. This explanation is consistent with the LSA cosine data, which showed that the textbook content deviated slightly from the canned text. The explanation is also consistent with the changes in computer literacy results we mentioned earlier in which reducing the textbook content to closely match the tutoring content increased its effectiveness (Graesser et al., 2003; Person, Graesser, Bautista, et al., 2001) and with other studies that have shown benefits for removing elaborations from textbooks (Charney, Reder, & Wells, 1988).

These three explanations are compatible with each other and may all have some truth. They offer explanations for why canned text remediation (from Experiment 1) may be more effective than reading a textbook (from Experiment 2 and earlier AutoTutor studies). However, we are still left with a counterintuitive finding: The interaction of the student with the computer tutors and the human tutors added no apparent value compared to the canned text remediation. This finding from Experiment 1 fails to support the widely believed interaction hypothesis. This motivated us to conduct a follow-up study to explore whether there are somewhat different circumstances or measures that might manifest differences between dialogue-based tutors and the canned text remediation condition.

5. Experiment 3: Improved assessments

It is conceivable that there really are differences in learning between the dialogue-based tutors and the canned text remediation conditions, but our assessments were not capable of detecting them in Experiment 1. Perhaps the tutored students learned deeper knowledge than the canned text remediation's students, but our tests measured only near-transfer problem solving under comparatively short retention intervals. This would be consistent with studies of AutoTutor in the computer literacy domain in which shallow test questions have shown no difference between AutoTutor and reading a textbook, but deep test questions have shown that AutoTutor students learned more than students reading a textbook (Graesser et al., 2003). This motivated us to repeat some of the conditions in Experiment 1 with potentially more sensitive assessments. In Experiment 3, we added a far-transfer essay test and a retention test 1 week later.

The tutored students of Experiment 1 spent twice or three times as long in training as the students in the canned text remediation condition, so they may have become fatigued or disengaged. The computer literacy experiments with AutoTutor had much smaller training times than those of Experiment 1. Therefore, in Experiment 3, we reduced the number of training problems to reduce the training time. In summary, in Experiment 3, we compared Why2-AutoTutor to the canned text remediation condition but used a far-transfer essay test, a 1-week retention test, and half as many training problems.

5.1. Method, materials, and participants

We reduced the number of training problems from 10 to 5. Some principles and misconceptions were no longer covered by training problems, so the tests were modified to cover only the

reduced set of principles and misconceptions. We created two similar versions of each test, A and B. We also wrote and administered far-transfer essay questions after a 1-week delay. We show the procedure used in Experiment 3 following.

Day 1
1. Pretest (3 essay questions).
2. Training (5 essay problems).
3. Near-transfer essay posttest (3 essay questions: Version A or B).
4. Multiple-choice posttest (26 questions: Version A or B).

Day 2 (retention testing)
1. Near-transfer essay posttest (3 essay questions: Version B or A).
2. Multiple-choice posttest (26 questions: Version B or A).
3. Far-transfer essay posttest (7 questions).

We reduced the pretest to a few essay questions because we were worried that more extensive testing might provide too much direction to the students' subsequent studying. The participants were student volunteers from the University of Memphis, Rhodes College, or Christian Brothers University who had already taken the relevant topics in a college physics class. There were 32 students in the Why2-AutoTutor condition and 30 students in the canned text remediation condition.

5.2. Results

The means and standard errors of the means are shown in Table 10 along with the p values obtained by ANOVAs and ANCOVAs. We scored the essays five ways: holistically (letter grades), principles counted with a stringent criterion, principles counted with a lenient criterion, misconceptions counted with a stringent criterion, and misconceptions counted with a lenient criterion. The adjusted test scores shown in Table 10 use the holistic pretest score as a covariate.

None of the posttest or adjusted posttest measures showed a reliable difference between the two conditions according to a two-tailed statistical test at $p < .05$. To save space, Table 10 omits two of the five kinds of essay scoring; the omitted measures were also not reliably different across conditions. Although none of the individual measures showed a significant difference, the measures collectively leaned toward an advantage for Why2-AutoTutor. Altogether, there were 17 adjusted posttest measures that compared Why2-AutoTutor and the canned text remediation condition. Why2-AutoTutor was highest for 14 measures, the canned text remediation condition was highest for 2 measures, and there was 1 tie. A Wilcox sign test showed a significant advantage for Why2-AutoTutor ($p < .05$). It would be appropriate to conclude that there was a very small but unreliable advantage for Why2-AutoTutor over the canned text remediation condition.

As in the earlier experiments, the tutored students took longer than the canned text remediation students to complete their training. The mean training time for the Why2-AutoTutor students (60.5 min; $SD = 15.1$) was significantly longer than the mean training time for the canned text remediation students (41.4 min; $SD = 20.5$), $F(1, 53) = 15.86$, $p < .001$, $MSE = 313.6$.

Table 10
Test score means (*M*) and standard errors (*SE*) in Experiment 3

Dependent Measure	Why2-AutoTutor		Canned Text Remediation		
	M	*SE*	*M*	*SE*	*p*
Pretest					
Holistic essay	.330	.040	.300	.040	.600
Stringent expectation	.136	.015	.114	.015	.285
Lenient misconception	.094	.014	.085	.015	.893
Posttest					
Immediate multiple choice	.725	.030	.650	.030	.084
Retention multiple choice	.734	.033	.661	.033	.122
Immediate essay (holistic)	.412	.043	.326	.043	.168
Retention essay (holistic)	.386	.046	.329	.046	.382
Immediate essay (stringent expectation)	.223	.025	.223	.025	.991
Retention essay (stringent expectation)	.217	.028	.198	.028	.632
Immediate essay (lenient misconception)	.056	.014	.058	.014	.912
Retention essay (lenient misconception)	.049	.011	.059	.011	.507
Far transfer (holistic)	.508	.043	.458	.043	.432
Far transfer (stringent expectation)	.275	.029	.213	.029	.142
Far transfer (lenient misconception)	.078	.013	.095	.013	.365
Adjusted posttests					
Immediate multiple choice	.720	.027	.660	.027	.097
Retention multiple choice	.730	.028	.670	.028	.136
Immediate essay (holistic)	.505	.045	.420	.045	.194
Retention essay (holistic)	.473	.050	.420	.050	.485
Immediate essay (stringent expectation)	.218	.025	.229	.025	.754
Retention essay (stringent expectation)	.209	.026	.207	.026	.972
Immediate essay (lenient misconception)	.056	.014	.058	.014	.910
Retention essay (lenient misconception)	.049	.010	.059	.010	.487
Far transfer (holistic)	.495	.033	.468	.033	.561
Far transfer (stringent expectation)	.265	.027	.222	.027	.268
Far Transfer (lenient misconception)	.077	.012	.095	.012	.312

Note.   *p* = two-tailed significance value.

To check for an aptitude–treatment interaction, we divided students into low- and high-prior-competence groups according to a median split on their holistic pretest scores. A 3× 2 ANOVA showed no significant Condition × Competence interaction for any of the posttest measures. As in the earlier experiments, there was no aptitude–treatment interaction for these participants who were preselected to have intermediate physics knowledge.

5.3. Discussion of Experiment 3

Despite the reduction in training and the improved assessments, in this experiment we showed the same results as in Experiment 1. The tutored students and the canned text remediation students learned about the same amount. There may have been a slight advan-

tage for Why2-AutoTutor, but the effect was small and unreliable. These results do not support the interaction hypothesis.

6. Experiment 4: Novice students

In all the preceding studies, the students had already taken physics or were currently taking physics and had taken the midterm on the topics that our studies focused on. These students had mastered some qualitative physics before the experiment. This is evident not only from the pretest scores (65% correct for Experiment 1) but from the comments during the dialogues in which they often mentioned concepts that would only be taught in physics courses. Moreover, physics textbooks written in recent years include many qualitative physics problems as well as the usual large number of quantitative problems. We do not know if our participants solved such problems when they took physics, but it is likely that at least some of them did. Thus, the participants in the first 3 studies should be considered intermediates rather than novices.

This observation suggests two reasons why the canned text remediation students learned as much as the tutored students. One reason is that when the intermediates were confused by the qualitative text they were reading, they could have used their quantitative knowledge to derive the missing qualitative knowledge. For instance, suppose the student did not immediately understand the canned text remediation when it said "Because the net horizontal force is zero, the horizontal acceleration is zero." An intermediate student could self-explain this statement using the quantitative version of Newton's Second Law, $F_X = m * a_x$. A novice who lacks knowledge of this equation might need a tutor's help to understand the assertion and learn the underlying qualitative principles.

A second reason why intermediates might have learned equally well from all the conditions is they may have been recovering access to old knowledge. Some intermediates had not taken physics recently, so their performance on the pretest was probably marred by recall failures. As they worked with the canned text remediation or any of the tutors, spreading activation and other memory effects may have increased their recall probability. By the posttest, they may have regained access to most of their original qualitative physics knowledge. Although this same reactivation could have occurred in the textbook condition of Experiment 2, it may be that the intermediates did not study the textbook carefully because they were not engaged in answering training questions. Thus, their memory access was at best only partially restored during Experiment 2's textbook condition and not enough to help them perform better than reading no physics at all.

These possibilities motivated us to compare the canned text remediation to human tutoring when the participants were novice students instead of intermediates. If prior knowledge of physics is required for effective learning from the canned text remediation, and if tutoring is truly effective, we would expect the novice students to learn more from human tutors than from the canned text remediation.

We also added a new condition wherein the student and the tutor communicated orally instead of via text. During the typed communication of Experiment 1, students had to wait an average of 24 sec per tutor turn for the tutor to finish typing. Such long pauses may have disrupted their learning. Typing their responses and the enforced turn taking may also have disrupted

their learning. Such disruptions could explain why the human tutors were not more effective than the canned text in Experiment 1. If so, then adding a spoken human tutoring condition should be more effective than the canned text remediation.

6.1. Method, material, and participants

The participants were paid volunteers from the University of Pittsburgh who had never taken college physics. Some had taken high school physics, and some had encountered physics concepts in nonphysics courses. For instance, math courses often use physics equations, and philosophy of science courses often deal with force in depth. Although the minds of these students were clearly not tabula rasa, they knew much less physics than the participants in our earlier studies.

The training problems and tests were identical to those used in Experiment 1 except that we did not administer the attitude test. The participants' knowledge of physics was spotty at best, so we developed a 10-page text based on Hewitt (1987). We removed all mathematical equations and only covered principles taught in the study. We presented principles in general terms and illustrated them with at most one example. Although we designed and pilot tested the canned text and the computer tutors with intermediates, we did not "water down" the canned text for the novices. The only change in materials was the addition of the 10-page pretraining text.

The procedure was the same as Experiment 1 with three exceptions. First, we only used one human tutor instead of four. Second, after the students took the pretest, they studied the 10-page textbook until they felt that they understood it completely. The mean time studying it was 32.0 min ($SD = 13.7$). They could not refer to the textbook as they worked with the canned text remediation or the human tutor. Third, the students worked with the human tutor for as many sessions as necessary to get though all 10 training problems. In a few cases, the students could not afford the time and energy to complete all 10 problems, so they skipped the last one ($N = 9$), two ($N = 4$), three ($N = 2$), or four ($N = 1$) problems for the typed human tutoring condition and the last one ($N = 8$) or two ($N = 1$) problems for the spoken human tutoring condition. Students in the canned text remediation condition always completed all 10 problems.

The spoken human tutoring condition was implemented by having the student and tutor in the same room separated by a partition. They viewed the same display as in the typed human tutoring (shown in Fig. 1), but they could not see each other. They could only hear each other. They used the shared display for essay entry only. In particular, the dialogue history window remained empty.

6.2. Results

Of the 25 students who started the typed human tutoring condition, 20 completed it. Of the 17 students who started the spoken human tutoring condition, 14 completed it. All 20 of the students who started the canned text remediation condition completed it. The multiple-choice pretest means of the students who dropped out of the experiment were identical to the pretest means of those who completed the experiment.

Table 11 presents the means and standard errors of all tests. Tests on the multiple-choice results revealed a statistically significant Condition × Test Phase interaction, $F(2, 51) = 9.82$, $p <$

Table 11
Test score means (*M*) and standard errors (*SE*) for Experiment 4

Dependent Measure	Human Spoken		Human Text		Canned Text Remediation		*p* hs > ht	*p* hs > ctr	*p* ht > ctr
	M	*SE*	*M*	*SE*	*M*	*SE*			
Pretest									
Multiple choice	.418	.027	.460	.022	.436	.022	.219	.619	.448
Essay (holistic)			.133	.031	.173	.037			.402
Essay (stringent expectation)			.054	.010	.052	.013			.888
Essay (lenient expectation)			.160	.027	.141	.030			.615
Essay (lenient misconception)			.047	.008	.040	.008			.539
Posttest									
Multiple choice	.727	.033	.671	.028	.565	.028	.186	.001	.013
Essay (holistic)			.615	.062	.475	.048			.080
Essay (stringent expectation)			.453	.054	.243	.054			.120
Essay (lenient expectation)			.612	.064	.538	.061			.402
Essay (lenient misconception)			.029	.009	.029	.008			.439
Adjusted posttest									
Multiple choice	.740	.030	.660	.025	.567	.025	.052	.000	.019
Essay (holistic)			.623	.055	.465	.055			.059
Essay (stringent expectation)			.543	.052	.337	.052			.127
Essay (lenient expectation)			.618	.061	.531	.061			.318
Essay (lenient misconception)			.029	.008	.020	.008			.450

Note.   hs = human tutoring spoken; ht = human tutoring typed; ctr = canned text remediation; *p* = two-tailed significance value.

.01, *MSE* = .007. The pattern of means clearly showed more learning gains from pretest to posttest in both human tutoring conditions than in the canned text remediation condition. An ANCOVA on the posttest scores using the pretest as the covariate showed statistically significant differences between conditions, $F(2, 50) = 10.27$, $p < .01$. The adjusted posttest scores showed the following ordering among means: human tutoring spoken > human tutoring typed > canned text remediation.

We conducted follow-up planned comparisons between pairs of conditions. None of the pairs of pretests were reliably different. All pairs of posttests were reliability different. Pairwise comparison of adjusted posttest scores showed typed human tutoring > canned text remediation ($p = .01$; effect size = 0.80), spoken human tutoring > canned text remediation ($p < .01$; effect size = 1.64), and spoken human tutoring > typed human tutoring ($p = .05$; effect size = 0.65). Assuming that all students covered the same inferences, these results support the interaction hypothesis.

To test for an aptitude–treatment interaction, we split the students into high-prior and low-prior competence via a median split on their multiple-choice pretest scores. A 3 × 2 ANOVA on posttest multiple-choice scores showed no significant Condition × Prior Competence interaction. Thus, there was no aptitude–treatment interaction in this experiment.

We scored the essay tests both by counting expectations and misconceptions and by assigning a holistic letter grade. Although there was a trend for the typed human tutoring students to

have better scores than the canned text remediation students, the differences were not reliable. As in Experiment 2, the essay tests appear to have been less sensitive to learning differences than the multiple-choice tests. Thus, we did not analyze the essays for the spoken human tutoring students.

We also performed analyses to examine possible differences in training time among the three conditions. We computed the time measures using two different theoretical considerations: total time spent during tutoring (total time) and time spent directly working with the tutor (work time). Using total time as the measure of student interaction, students in the typed human tutoring condition ($M = 440.8$, $SD = 170.8$) took much more time to finish their training than students in the spoken human tutoring condition ($M = 166.6$, $SD = 45.1$) and the canned text remediation ($M = 85.4$, $SD = 38.4$) conditions, $F(2, 48) = 56.9$, $p < .001$; differences were not significant in the latter two conditions. The second measure of student time involvement, work time, excludes the time that students spent waiting for the tutors to reply in the typed human tutoring condition. An ANOVA on work time indicated that the typed human tutoring condition ($M = 208.5$, $SD = 92.9$) took significantly more time than in the spoken human tutoring ($M = 166.6$, $SD = 45.1$) and the canned text remediation ($M = 85.4$, $SD = 38.4$) conditions, $F(2, 48) = 18.1$, $p < .001$. Because the work time data showed typed human tutoring > canned text remediation = spoken human tutoring, one can argue that time on task explains the superiority of typed human tutoring over canned text remediation, but one cannot argue that time on task explains the overall superiority of spoken human tutoring.

7. Experiment 5: All four conditions with novice students

Because Experiment 4 showed that novices learned more with a human tutor than with the canned text remediation, we felt confident that we could repeat Experiment 1 with novices students and see the hypothesized pattern of human tutoring > canned text remediation, and moreover, we could compare the effectiveness of the computer tutors to the human tutors. However, to reduce the number of conditions and the training time, we used only the spoken human tutoring condition and not the typed human tutoring condition. Moreover, so that the entire experiment could be completed in a single session, we used the abbreviated training materials of Experiment 3 in this experiment.

7.1. Methods, materials, and participants

The participants were volunteers who had not taken a physics course in college. We drew them from the University of Pittsburgh, the University of Memphis, and Rhodes College. We randomly assigned them to a human tutors ($N = 21$), Why2-AutoTutor ($N = 21$), Why2-Atlas ($N = 23$), and the canned text remediation ($N = 19$). All students who started the experiment completed it.

The procedure was somewhat different than earlier experiments. Although we used the abbreviated training materials of Experiment 3, we did not include a retention test. Unlike Experiment 4, the pretest occurred after the booklet was read instead of before it. This allowed us to measure gains caused by the manipulations themselves and without confusing them with gains

caused by reading the booklet. Unlike Experiment 4, we used two tutors, one in Pittsburgh and one in Memphis. The students and the human tutors conversed via the telephone (using a headset). This allowed us to use Memphis tutor with Pittsburgh students and vice versa. The procedure was

1. Study the 10-page textbook.
2. Pretest (three essay questions; 26 multiple-choice questions; Test A or B).
3. Training (five essay problems).
4. Near-transfer posttest (three essay questions; 26 multiple-choice questions; Test B or A).
5. Far-transfer posttest (seven essay questions).

We slightly modified Why2-Atlas and the canned text remediation. Why2-Atlas added for each problem a "walk through" KCD that elicited the whole line of reasoning for answering the problem at a somewhat abstract level of detail. Why2-Atlas also included more misconceptions in its essay analyzer and the corresponding misconception KCD. To maintain content equivalence between Why2-Atlas and the canned text remediation, we rewrote the canned text remediation minilessons to include monologue versions of the new KCD.

7.2. Results

We discuss the multiple-choice tests first. Table 12 shows the means and standard errors for all four conditions. The pretest scores were not reliably different among the four conditions, $F(3, 80) = 1.32$, $p = .27$, $MSE = .043$. In the ANOVA with the Condition × Test Phase factorial design, there was a robust main effect for Test Phase, $F(1, 80) = 154.50$, $p < .001$, $MSE = 0.009$, but there was no significant main effect of experimental condition and no significant Condition × Test Phase interaction. An ANCOVA with pretest scores as the covariate showed that the adjusted posttest scores of the conditions were not reliably different overall, $F(3, 79) = 1.20$, $p =$

Table 12
Test score means (*M*) and standard errors (*SE*) for Experiment 5

Dependent Measure	Human Spoken		Why2-Atlas		Why2-AutoTutor		Canned Text Remediation	
	M	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*
Pretest								
Multiple choice	.485	.039	.495	.038	.418	.039	.407	.041
Essay (holistic)	.310	.047	.304	.045	.230	.047	.176	.049
Posttest								
Multiple choice	.683	.041	.697	.039	.599	.041	.563	.043
Near-transfer essay (holistic)	.623	.064	.536	.061	.504	.064	.426	.067
Far-transfer essay (holistic)	.416	.051	.531	.048	.455	.050	.393	.051
Adjusted posttest								
Multiple choice	.659	.028	.666	.027	.626	.028	.597	.030
Near-transfer essay (holistic)	.585	.054	.502	.052	.525	.054	.487	.058
Far-transfer essay (holistic)	.377	.041	.499	.038	.472	.040	.452	.042

.32, $MSE = 0.020$, nor were there significant pairwise differences. In summary, the students in all four groups learned, and they learned about the same amount. This is exactly the same pattern of results as seen in Experiment 1.

The essay tests showed the same pattern. Table 12 shows the means and standard errors for all four conditions. The pretest scores were not reliably different among the four conditions, $F(3, 80) = 1.81$, $p = .15$, $MSE = 0.740$. For the near-transfer posttests, in the ANOVA with Condition × Test Phase factorial design, there was a significant main effect for Test Phase, $F(1, 80) = 94.80$, $p < .001$, $MSE = .504$, but no significant main effect of experimental condition and no significant Condition × Test Phase interaction. An ANCOVA with pretest scores as the covariate showed that the adjusted near-transfer posttest scores of the conditions were not reliably different overall, $F(3, 75) = 0.26$, $p = .853$, $MSE = 0.982$, nor were there significant pairwise differences. The far-transfer posttests were missing for four students (two in the human tutoring condition, one in the Why2-Atlas condition, and one in the Why2-AutoTutor condition) due to experimenter error. In the ANOVA with Condition × Test Phase factorial design, there was a significant main effect for Test Phase $F(1, 76) = 48.56$, $p < .001$, $MSE = 0.492$, but no significant main effect of experimental condition and no significant Condition × Test Phase interaction. An ANCOVA with pretest scores as the covariate showed that the adjusted far-transfer posttest scores of the conditions were not reliably different overall, $F(3, 75) = 1.72$, $p = .17$, $MSE = .502$. Thus, the essay results duplicated the multiple-choice results in showing that students learned in all four conditions, and they learned about the same amount.

To check for an aptitude–treatment interaction, we divided students into high- and low-prior-competence groups using a median split on their multiple-choice pretest scores. In a 3 × 2 ANOVA on the multiple-choice posttest scores, there was a marginal Condition × Competence interaction, $F(3, 76) = 2.24$, $p = .09$, $MSE = .024$. An ANCOVA, with pretest scores as the covariate, of the high-pretest group showed that the adjusted posttest scores were not reliably different across conditions. However, an ANCOVA of the low-pretest group showed that their adjusted posttests scores were different, $F(3, 32) = 3.357$, $p = .031$, $MSE = .021$. Table 13 shows the means and standard errors for the low-pretest group. Pairwise comparisons of the adjusted multiple-choice posttest scores show that spoken human tutoring > canned text remediation ($p = .017$; effect size = 0.69) and spoken human tutoring > Why2-AutoTutor ($p = .021$; effect size = 0.37), but none of the other comparisons were reliable. Pairwise comparisons of the adjusted near-transfer essay test scores showed a similar pattern as the multiple-choice data, namely, that spoken human tutoring > Why2-AutoTutor ($p = .038$) and spoken human tutoring > Why2-Atlas ($p = .01$). The comparison of spoken human tutoring with canned text remediation was only marginally significant ($p = .075$). For the adjusted far-transfer essay test scores, none of the pairwise comparisons was reliable. Thus, it appears that for the low-pretest students, spoken human tutoring was more effective than canned text remediation.

To summarize, when all students are considered, the multiple-choice tests, near-transfer essay tests, and far-transfer essay tests all showed a null result when comparing the different tutoring conditions. Although students in all four conditions learned considerably, they learned about the same amount. On the other hand, when considering only the students with low pretest scores, the human tutees learned significantly more than the canned text remediation students. The relation between the computer tutors and the other conditions was less clear, but it

Table 13
Test score means (*M*) and standard errors (*SE*) for Experiment 5 using low-pretest students

Dependent Measure	Human Spoken[a]		Why2-Atlas[b]		Why2-AutoTutor[c]		Canned Text Remediation[d]	
	M	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*
Pretest								
Multiple choice	.321	.024	.286	.027	.277	.023	.301	.022
Essay (holistic)	.178	.046	.167	.052	.141	.043	.115	.041
Posttest								
Multiple-choice	.645	.053	.582	.060	.438	.050	.462	.048
Near-transfer essay (holistic)	.444	.068	.178	.078	.230	.065	.242	.062
Far-transfer essay (holistic)	.267	.051	.224	.062	.283	.050	.252	.046
Adjusted posttest								
Multiple choice	.622	.049	.593	.055	.457	.046	.457	.043
Near-transfer essay (holistic)	.444	.060	.179	.073	.229	.060	.242	.054
Far-transfer essay (holistic)	.267	.044	.224	.054	.283	.055	.252	.040

[a]*n* = 9. [b]*n* = 7. [c]*n* = 10. [d]*n* = 11.

appears that their students' learning gains with low-pretest students may have fallen in between those of spoken human tutoring and canned text remediation.

7.3. Discussion of Experiment 5

Experiments 4 and 5 were motivated by the hypothesis that interactive instruction and noninteractive instruction are equally effective for relearning (as opposed to learning), whereas interactive instruction shows advantages for learning. To test the hypothesis when students were learning (as opposed to relearning), in Experiments 4 and 5, we used students who had not taken college physics. The interaction hypothesis predicts that in both experiments, the tutees should have learned more than the readers who merely worked with canned text remediation. The results of Experiment 4 were consistent with this prediction, but only the low-pretest students in Experiment 5 fared better with tutoring than canned text remediation. The aptitude–treatment interaction of Experiment 5 is not completely mysterious because it has often been found that high-aptitude students can learn from any instruction, whereas low-aptitude students learn more when the instruction uses more scaffolding (Cronbach & Snow, 1977).

Nonetheless, in looking over the transcripts from the tutoring conditions of Experiments 4 and 5, we noticed that the novices seemed to have had a great deal of trouble understanding the material even with the aid of a tutor. This makes sense because the material was designed for students who had already completed a college physics course, and these students had not. Moreover, the students were not allowed to refer back to the 10-page textbook that they studied during pretraining. These conditions seemed atypical of real-world learning, so we explored whether the interaction hypothesis persisted when novices learned from material that was written for them.

8. Experiments 6 and 7

Although the interaction hypothesis was supported by Experiment 4 and partially supported by Experiment 5, the instructional content of those experiments was aimed at intermediates rather than the novice students used in those experiments. To test the interaction hypothesis in a more realistic learning situation, we again used novices so that the participants would be learning and not relearning, but we modified the instruction to make it appropriate for novices. In particular, we modified the training's content to make it easier to learn, and we provided extensive pretraining so that all the students would have strong enough prior knowledge.

We also added a new condition to tease apart two explanations offered for Experiment 2's finding that Why2-AutoTutor was more effective than just reading a textbook. Because Why2-AutoTutor tied with the canned text remediation in other experiments, Experiment 2 implied that canned text remediation was more effective than the textbook. Two explanations were offered:

- The canned text had higher coherence and slightly different content than the textbook. In particular, the canned text focused on solving problems, whereas the textbook focused on introducing principles and concepts.
- The canned text remediation required students to answer the training questions, whereas the textbook neither posed questions nor asked students to write anything.

To differentiate these explanations, in this experiment, we included a text only condition wherein the students studied the canned text, including the questions and their ideal answers, but did not write answers of their own. That is, students in the text only condition only read text and never answered questions. If the textbook was ineffective because of its content and coherence, then the text only condition should tie with the canned text remediation. If the textbook condition was ineffective because it did not require students to answer questions, then the text only conditions should be less effective than the canned text remediation. Moreover, if we make the plausible assumption that the text only condition is less interactive than the canned text remediation condition, then the interaction hypothesis predicts that the text only condition should be less effective than the canned text remediation condition.

Last, we dropped the human tutoring condition because we doubted that its content could be equated with the others. Although our pretraining would teach the prerequisites, we could not be certain that all the students would master them, in which case the tutors would probably teach the prerequisites, whereas our computer tutors and the canned text remediation would not. To insure that all the conditions covered the same content, we used only the computer tutors.

The interaction hypothesis predicts that the four conditions should be ordered by effectiveness as

Text only < canned text remediation < (Why2-AutoTutor, Why2-Atlas)

In particular, we had no prediction about which of the two computer tutors would be more effective.

8.1. Method, materials, and participants

We used the same participants and methods as Experiment 5, but we changed the materials significantly. The old materials were intended for students who had taken college physics, whereas these materials were designed for students who had never taken physics.

First, we reduced the set of principles to be taught. We selected four of the training problems used in an earlier experiment. They are problems 1, 3, 5, and 9 in Appendix A. Of the many solutions available for each problem, we selected just one so that we could reduce the number of principles to be taught. For instance, the intermediates often knew that "all objects in free-fall have the same acceleration, g." This is not a fundamental principle of physics, but it can be derived from Newton's law and the weight law, which are more fundamental principles of physics. We wanted to simplify the learning task, so we targeted only Newton's law and the weight law and did not teach the free-fall "theorem." Similarly, there are six kinematics equations for constant acceleration that are familiar to intermediates, but three suffice for deriving the others, so we taught only those three.

Second, we specified precise instructional objectives. They were

- To understand the difference between a vector and a scalar.
- To be able to distinguish vector principles (e.g., Newton's law) from scalar principles (e.g., the weight law).
- To be able to state each principle in a generic form and to enclose vector principles in angle brackets, for example, "<net force = mass * acceleration>" and "<net force = sum of individual forces>.
- To be able to qualitatively apply each principle in isolation. For instance, the following is a problem that can be solved by qualitative application of one principle, the definition of average velocity:

  A rabbit and a turtle have a race. During the first minutes of the race, the turtle plods about 5 feet. Meanwhile, the rabbit dashes to the finish line, then scampers back to the starting line, where it shouts, "I am so much faster than you! I ran the race twice, and you've only gone a few feet." The turtle replies, "Yes, but my average velocity is greater." Explain.
- To be able to qualitatively apply several principles in combination by using a forces acceleration velocities displacement (FAVD) strategy. There are principles that connect each quantity in the sequence to the next quantity in the sequence. For instance, Newton's second law connects forces to acceleration, and the definition of average velocity connects velocity to displacement. Students should learn to start at the given quantity and apply principles in sequence until they arrive at the sought quantity.

Some of these concepts and strategies are not taught in physics courses. Although we believe the additions simplify student's learning of this particular content, it is possible that they are inconsistent with instructional objectives of physics courses. Thus, we are not advocating their inclusion in physics courses, but they served our purposes in Experiments 6 and 7.

Third, we decided which objectives would be taught during the training, where the manipulation would occur, and which would be taught during the pretraining. The last objective in the preceding list was taught during training, whereas all the others were taught during pretraining.

Fourth, we developed the pretraining, which consisted of seven lessons, one for each major principle. Each lesson consisted of about two pages of text and one or more canned text remediation exercises. For instance, the rabbit–turtle problem we quoted previously was part of the canned text remediation for the average velocity lesson.

Fifth, we modified the training. We modified Why2-Atlas and Why2-AutoTutor so that they both taught all and only the instructional objectives we listed previously. For instance, they used angle brackets around vector principles and explicitly mentioned FAVD. Although we changed the dialogue management and natural language understanding of Why2-AutoTutor very little, we changed Why2-Atlas extensively (Jordan et al., in press). For each of the four training problems, canned text remediation was written to have the same content as Why2-Atlas. The text only materials were identical to the canned text remediation, but the text only software did not ask the students to enter essays.

Sixth, we developed a new pretest and a new posttest. The new pretest had 15 multiple choice and two essay questions. The new posttest had 14 multiple-choice, five fill-in-the-blank, and six essay questions. Each fill-in-the-blank problem asked a top level question that was similar in complexity to the essay questions but provided a paragraph-long answer with blanks in key places. That is, the fill-in-the-blanks problems assessed students' ability to compose a multi principle explanation given some scaffolding.

The resulting materials were more focused and coherent than those used in the earlier experiments. They addressed a smaller set of principles and concepts. The prerequisites of each had been identified and taught in pretraining. The test items assessed each principle and concept in multiple contexts.

As in the earlier experiments, paid volunteers were drawn from students at the University of Pittsburgh, the University of Memphis, and Rhodes College. To insure that only novices participated, the advertisements did not mention this requirement. Instead, we asked students about their physics background during the initial telephone contact, and we accepted only those with no college physics.

Although we ran all four conditions during Experiment 6, the data from the Why2-Atlas students was corrupted due to a software error. This was only discovered during data analysis. Thus, we ran another experiment using only two conditions: Why2-Atlas and the canned text remediation. Because Experiments 6 and 7 used the same materials and drew participants from the same populations, albeit at different times, we discuss their results together.

8.2. Results

The pretest scores of Experiment 6 and 7 were not significantly different overall, nor were the posttest scores for the canned text remediation conditions of the two experiments. Thus, we pooled the data from the two conditions. Table 14 shows the means and standard errors of the pretests scores, posttests scores, and posttests scores adjusted by the ANCOVAs we discuss following.

The pretest scores were not reliably different among the four conditions: multiple choice, $F(3, 163) = 0.99$, $p = .395$, $MSE = 0.31$; essay, $F(3, 163) = 1.076$, $p = .361$, $MSE = .036$. The posttest scores were not reliability different overall: multiple choice, $F(3, 163) = 0.70$, $p = .557$, $MSE = 0.017$; essay, $F(3, 163) = 0.69$, $p = .573$, $MSE = .022$; fill in the blanks, $F(3, 163) = 2.47$, $p = .064$, $MSE = .080$. Pairwise comparisons of the posttests scored showed only one significant differ-

Table 14
Test score means (*M*) and standard errors (*SE*) for Experiments 6 and 7

Dependent Measure	Why2-Atlas[a]		Why2-AutoTutor[b]		Canned Text Remediation[c]		Text Only[d]	
	M	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*
Pretest								
Multiple choice	.550	.028	.516	.034	.511	.021	.477	.032
Essay (holistic)	.204	.029	.213	.035	.180	.022	.137	.033
Posttest								
Multiple choice	.520	.025	.472	.030	.487	.019	.514	.028
Essay (holistic)	.365	.029	.329	.035	.324	.022	.366	.033
Fill in the blank	.546	.029	.455	.035	.452	.022	.481	.032
Adjusted posttest								
Multiple choice	.503	.022	.472	.026	.489	.016	.531	.024
Essay (holistic)	.354	.028	.328	.033	.325	.021	.377	.031
Fill in the blank	.539	.028	.455	.034	.453	.021	.489	.032

[a]*n* = 39. [b]*n* = 27. [c]*n* = 70. [d]*n* = 31.

ence: The Why2-Atlas students scored higher on the fill-in-the-blanks test than the canned text remediation students, *F*(1, 74) = 6.33, *p* = .010. We ran three ANCOVAs using the multiple-choice pretest scores as a covariate, and none showed significant differences among the four conditions: multiple-choice, *F*(3, 161) = 1.10, *p* = .353, *MSE* = .019; essay, *F*(3, 161) = 0.77, *p* = .514, *MSE* = .023; fill in the blanks, *F*(3, 161) = 2.16, *p* = .095, *MSE* = .095. Only one pairwise comparison was significant: The Why2-Atlas students had higher adjusted posttest scores on the fill-in-the-blanks test than the canned text remediation students, *F*(1, 72) = 5.83, *p* = .018. Although this reliable difference was in the direction predicted by the interaction hypothesis, it was not accompanied by similar differences for the multiple-choice test and the essay test.

To check for aptitude–treatment interaction, we divided students into high- and low-prior-competence groups using a median split on the multiple-choice pretest scores. A 4 × 2 ANOVA on each of the posttest scores showed no significant Condition × Prior Competence interaction. Thus, there was no aptitude–treatment interaction in this experiment.

Except for one pairwise comparison in which Why2-Atlas students scored higher on the fill-in-the-blanks posttest than the canned text remediation students, the pattern of results was the same as Experiment 1 and inconsistent with the interaction hypothesis: All students learned the same amount. Moreover, it appears that the text only condition is just as effective as the canned text remediation condition, which suggests that in this experiment, having students answer essay questions did not increase the effectiveness of the instruction compared to simply reading the canned text.

9. General discussion

Table 15 summarizes the results. For simplicity, it collapses the four kinds of tutoring. That is, the spoken human tutoring, typed human tutoring, Why2-Atlas, and Why2-AutoTutor con-

Table 15
Summary of experimental results

Student	Content	Experiment	Result
Intermediate	Intermediate	1 and 3	Tutoring = canned text remediation
Novice	Novice	6 and 7	Tutoring = canned text remediation = text only
Intermediate	?	2	Tutoring > textbook = no instruction
Novice	Intermediate	4 and 5?	Tutoring > canned text remediation

ditions are all represented in Table 15 as "Tutoring." As the first row indicates, Experiments 1 and 3 involved intermediate students learning intermediate-level content, and tutoring was no more effective than canned text remediation. As the second row shows, Experiments 6 and 7 involved novices learning novice-level material, and again tutoring was no more effective than canned text remediation. However, as shown in the third row, tutoring was more effective than a textbook in Experiment 2, which involved intermediate students. However, it is not clear how to classify the content of Experiment 2 because the textbook may not have covered the same content as the tutoring. Finally, the fourth row indicates that when novice students were taught intermediate content, tutoring was more effective than canned text remediation. This occurred in Experiment 4 and with the low-pretest students of Experiment 5.

The interaction hypothesis predicts that tutoring should always be better than the low-interaction, reading-based control conditions. This prediction was not supported when the level of the content was the same as the level of the students. That is, when intermediate students studied material written for intermediates (Experiments 1 and 3) or novice students studied material written for novices (Experiments 6 and 7), then tutoring was no more effective than the low-interaction control conditions.

However, when intermediate students studied the textbook and were not required to answer questions during training (Experiment 2), then tutoring was more effective than reading. Moreover, reading the textbook was not effective at all because it tied with a test–retest control condition in which students received no training. Although these results were consistent with the interaction hypothesis, they also suggest that the intermediates processed the text very shallowly and/or that the textbook's content was not as aligned with the posttests as the tutoring's content.

As shown in the last row of Table 15, when novices studied material written for intermediates, the interaction hypothesis was partially supported because tutoring is sometimes more effective than the canned text remediation. However, we still need to explain why the interaction hypothesis was not supported by the high-pretest students of Experiment 5 and why there was an aptitude–treatment interaction on Experiment 5 but not on any other experiment.

An explanation can be formulated using a familiar concept in developmental psychology, the zone of proximal development (ZED). Branford, Brown, and Chocking (2000) defined the ZPD as "the distance between the actual developmental level as determined by independent problem solving and the level of potential development as determined through problem solving under adult guidance, or in collaboration with more capable peers" (p. 81). This concept is illustrated by Fig. 3. The lower, light gray parts of the bars represent the range of complexity that students
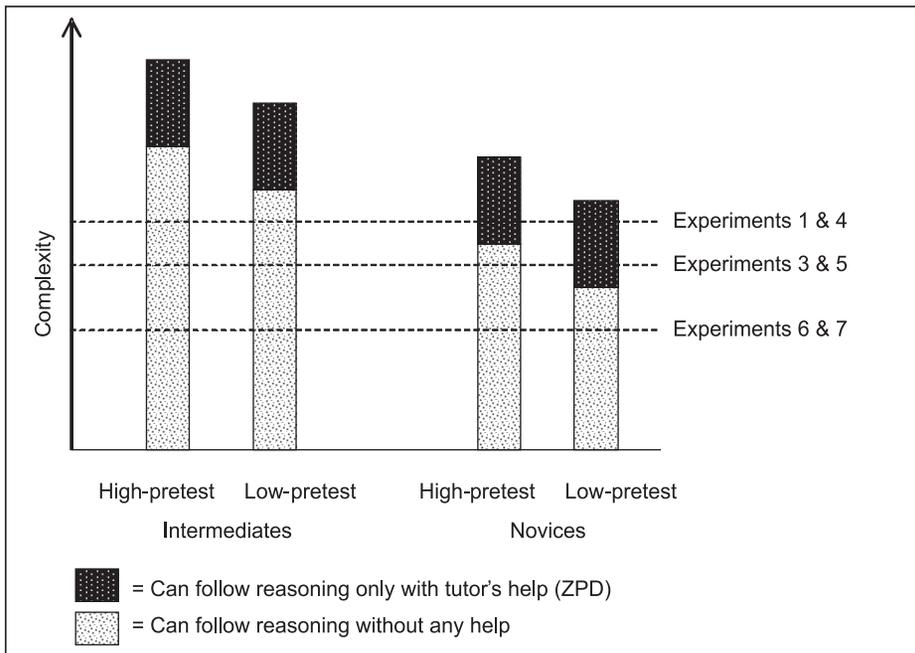
Fig. 3. Hypothesized zones of proximal development (ZPD).

handle independently without the aid of a tutor. The upper, dark gray part of the bars represents the range of complexity that students can handle given support from a tutor. That is, the upper part of the bars represents the ZPD. As Fig. 3 illustrates, we can plausibly assume that our intermediates could handle more complexity without help than the novices and that high-pretest students could handle more complexity without help than the low-pretest students.

The dotted horizontal lines in Fig. 3 represent the complexity of the content in the experiments. Experiments 1 and 4 used the same materials, which were designed for intermediates. Although the materials of Experiments 3 and 5 were also written for intermediates, they were simpler than the materials used in Experiments 1 and 4 because they had half as many training problems and covered fewer physics concepts. Thus, the line representing their complexity is lower. The lowest line represents the complexity of the materials used in Experiments 6 and 7, which were designed for novices.

We have not included Experiment 2 in Fig. 3. Although its tutoring condition used the same materials as Experiments 1 and 4, it is not clear where to place the level of complexity of the textbook.

Given these assumptions, we can explain all the results of these experiments including the one-time-only appearance of an aptitude–treatment interaction. We discuss the experiments in the order in which they appear in Fig. 3.

- In Experiment 1, we used intermediates, and the complexity of the materials fell within the range of complexity that they could handle without help; so tutoring provided no advantage to them, and there was no aptitude–treatment interaction.

- In Experiment 4, we used novices, and the materials fell into the ZPD of both high-pretest and low-pretest students. Thus, tutoring was required for the materials to be effective, and there was no aptitude–treatment interaction.
- In Experiment 3, we used intermediates, and the materials were appropriate for them to handle independently, so there was no advantage for tutoring and no aptitude–treatment interaction.
- In Experiment 5, we used novices. The materials could be handled independently by the high-pretest students, so tutoring provided no advantage to them. The content was in the ZPD of the low-pretest students, so tutoring did provide an advantage to them. Thus, there was an aptitude–treatment interaction.
- In Experiments 6 and 7, we used novices, and we designed their materials to be comprehensible by most students even when studying alone. Thus, tutoring had no advantage, and there was no aptitude–treatment interaction.
- In Experiment 2, we used intermediates. As we discussed earlier, the advantage of tutoring could be due to differences in the content of the textbook versus the tutoring or to shallow processing of the textbook due to the absence of questioning.

In short, these data indicate that if students are given instructional content that is designed for students at their level of preparation, then tutoring has no advantage over studying text alone. However, if the content is in the students' ZED, then tutoring has a big advantage over studying text alone. Indeed, the effect size in Experiment 4 was an impressive 1.64 standard deviations for spoken human tutoring versus canned text remediation.

9.1. Future work

There are several important opportunities for future work. For starters, the generality of our results should be tested. Our experiments have all been conducted in a laboratory setting with college students and a modest amount of material in qualitative physics. Perhaps canned text remediation and tutoring would fare differently if they were compared over a longer period of instruction in a more realistic setting such as a LearnLab course (a set of real classes that have been instrumented for fine-grained data collection—see www.learnlab.org).

Second, our results must be reconciled with those of the earlier experiments testing the interaction hypothesis. As we pointed out in the literature review, control conditions that were completely passive tended to fare worse than tutoring, whereas control conditions that required the student to answer questions or solve problems during training tended to be just as effective as tutoring. This hypothesis predicts that the text only condition of Experiment 6, which was completely passive, should have produced smaller gains than tutoring; but it did not. It also predicts that the canned text remediation conditions of Experiment 4, which did have questions embedded in it, should have produced the same gains as tutoring; yet it produced smaller ones. Thus, we need to reexamine the earlier studies to see if support for this hypothesis evaporates once we take into consideration the control of content and the ZPD. VanLehn (in press) provided one such reconciliation.

Third, although we have found an explanation that is consistent with the results from all seven experiments, it would be beneficial to have a more detailed explanation. As examples,

in the next few paragraphs, we sketch three detailed potential explanations of our main result.

9.1.1. Content differences

Despite our attempts to equate content across conditions, human tutors working with tutees who were having difficulties understanding the content may have provided crucial extra content. In particular, tutors may have provided instruction on presuppositions of the text that were not familiar to the students. For instance, our intermediate-level canned text presupposed that readers were familiar with the idea of analyzing motion by separately analyzing its horizontal and vertical components. For students who have taken a semester of college physics, this presupposition is thoroughly familiar. For novices, it may not have been familiar, which may have made the text difficult to follow. Yet, with tutors supplying this crucial extra content, perhaps the novices could follow the explanations and learn from them. This would explain why content that was "in the student's ZPD" (they could follow the reasoning only with the aid of a tutor) elicited more learning from tutees than readers.

9.1.2. Engagement differences

Even if the content of the tutoring were completely equivalent to the content of the text, the tutees may have paid more attention to that content than the readers. When following a complex line of reasoning, a tutee must continually respond to the tutor's questions, whereas a reader's attention could wander and miss some steps along the line of reasoning. If the material is so difficult that students become frustrated, then readers may deliberately skim the text, whereas tutees who deliberately disengage risk offending a human tutor. When more steps in the lines of reasoning are skipped (deliberately or accidentally) by readers than tutees, then the explanation is in the student's ZPD and elicits more learning from tutees than readers. Thus, the step-skipping hypothesis explains our main result.

9.1.3. Cognitive differences

Even if both the content and the engagement of students were completely equivalent so that both tutees and readers processed exactly the same steps in the same lines of reasoning, there may be advantages in eliciting a step compared to presenting it. Compare, for instance, a presentation of a step with two tutorial elicitation of the same step:

A. Text: The force on the car due to earth's gravity is called the weight force on the car.
B. Tutor: What is the force on the car due to the earth called?
   Student: The weight force on the car.
C. Tutor: What is the force on the car due to the earth called?
   Student: Gravitational.
   Tutor: Yes, but it is also called the weight force on the car.

Paired-associate learning effects, such as the self-generation effect (Slamecka & Graf, 1978), suggest that successful elicitation episodes, such as the preceding B, should be more effective than presentations such as the preceding A. This seems inconsistent with our data, as it would predict an advantage for tutoring even when students are not in their ZPD. However, when students are in their ZPD, they cannot completely follow a line of reasoning without the tutors'

help, so the tutorial dialogue will have frequent elicitation episodes with negative or partially negative feedback such as the preceding C. If tutees received explicit feedback as in Episode C learned more than readers who diligently self-explained a presentation, such as Episode A (Chi, Basso, Lewis, Riemann, & Glaser, 1989; Renkl, 2002), then this step-level hypothesis explains our instruction-level observation that tutees learn more than readers when working with material in their ZPD but not when working with less complex content.

These are just three explanations of our main result. There are probably many others. An important next step in the research would be studies that determine which explanations are most accurate. Determining when and why tutoring is more effective than less interactive instruction is important not only for helping to understand student cognition, but also, in the long run, for helping to develop better educational technology and more cost-effective educational policies.

Notes

1. The United States program called "No Child Left Behind" uses *tutoring* to include an instructor working with a small group of students, whereas we use *tutoring* to refer only to one-on-one instruction.
2. Note that the tutor aligned the tutoring with the research methods course curriculum. The material on *t* tests at that point in the curriculum addressed comparisons between two conditions but not comparisons of a sample mean to a fixed value.

Acknowledgments

References

Aleven, V., Ogden, A., Popescu, O., Torrey, C., & Koedinger, K. (2004). Evaluating the effectiveness of a tutorial dialogue system for self-explanation. In J. Lester, R. M. Vicari, & F. Paraguaca (Eds.), *Intelligent Tutoring Systems: Seventh International Conference, ITS 2004* (pp. 443–454). Berlin, Germany: Springer.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (2000). *How people learn: Brain, mind, experience and school.* Washington, DC: National Academy Press.

Charney, D. H., Reder, L. M., & Wells, G. W. (1988). Studies of elaboration in instructional texts. In S. Doheny-Farina (Ed.), *Effective documentation: What we have learned from research* (pp. 47–72). Cambridge, MA: MIT Press.

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, 15,* 145–182.

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

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.

Cohen, P. A., Kulik, J. A., & Kulik, C.-L. C. (1982). Educational outcomes of tutoring: A meta-analysis of findings. *American Educational Research Journal, 19,* 237–248.

Collins, A., Brown, J. S., & Newman, S. E. (1989). Cognitive apprenticeship: Teaching the craft of reading, writing and mathematics. In L. B. Resnick (Ed.), *Knowing, learning and instruction: Essays in honor of Robert Glaser* (pp. 453–494). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Core, M. G., Moore, J. D., & Zinn, C. (2003). The role of initiative in tutorial dialogue. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL)* (pp. 67–74). Morristown, NJ: Association of Computation Linguistics.

Craig, S. D., Driscoll, D. M., & Gholson, B. (2004). Constructing knowledge from dialog in an intelligent tutoring system: Interactive learning, vicarious learning and pedagogical agents. *Journal of Educational Multimedia and Hypermedia, 13,* 163–183.

Craig, S. D., Sullins, J., Witherspoon, A., & Gholson, B. (in press). The deep-level reasoning questions effect: The role of dialog and deep-level reasoning questions during vicarious learning. *Cognition and Instruction.*

Cronbach, L. J., & Snow, R. E. (1977). *Aptitudes and instructional methods: A handbook for research on interactions.* New York: Irvington.

de Kleer, J. (1977). Multiple representations of knowledge in a mechanics problem-solver. In *International joint conference on Artificial Intelligence* (pp. 299–304). Cambridge, MA: MIT Press.

di Sessa, A. A. (1993). Towards an epistemology of physics. *Cognition and Instruction, 10,* 105–225.

Evens, M., & Michael, J. (2000). *One-on-one tutoring by humans and machines.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Foltz, P. W., Gilliam, S., & Kendall, S. (2000). Supporting content-based feedback in on-line writing evaluation with LSA. *Interactive Learning Environments, 8,* 111–127.

Freedman, R., Rosé, C. P., Ringenberg, M. A., & VanLehn, K. (2000). ITS Tools for natural language dialogue: A domain-independent parser and planner. In G. Gauthier, C. Frasson, & K. VanLehn (Eds.), *Intelligent Tutoring Systems: 5th International Conference, ITS 2000* (pp. 433–442). Berlin, Germany: Springer.

Graesser, A. C., Lu, S., Jackson, G. T., Mitchell, H. H., Ventura, M., Olney, A., et al. (2004). AutoTutor: A tutor with dialogue in natural language. *Behavioral Reseach Methods, Instruments and Computers, 36,* 180–193.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohension and language. *Behavioral Reseach Methods, Instruments and Computers, 36,* 193–202.

Graesser, A. C., Moreno, K., Marineau, J., Adcock, A., Olney, A., & Person, N. (2003). AutoTutor improves deep learning of computer literacy: Is it the dialog or the talking head? In U. Hoppe, F. Verdejo, & J. Kay (Eds.), *Proceedings of Artificial Intelligence in Education* (pp. 47–54). Amsterdam: IOS.

Graesser, A. C., & Olde, B. A. (2003). How does one know whether a person understands a device? The quality of questions the person asks when the device breaks down. *Journal of Educational Psychology, 95,* 524–536.

Graesser, A. C., & Person, N. (1994). Question asking during tutoring. *American Educational Research Journal, 31,* 104–137.

Graesser, A. C., Person, N., Harter, D., & Tutoring Research Group. (2001). Teaching tactics and dialog in AutoTutor. *International Journal of Artificial Intelligence and Education, 12,* 257–279.

Graesser, A. C., Person, N., & Magliano, J. (1995). Collaborative dialog patterns in naturalistic one-on-one tutoring. *Applied Cognitive Psychology, 9,* 359–387.

Graesser, A. C., Wiemer-Hastings, K., Wiemer-Hastings, P., & Kreuz, R. (1999). AutoTutor: A simulation of a human tutor. *Journal of Cognitive Systems Research, 1,* 15–25.

Graesser, A. C., Wiemer-Hastings, P., Wiemer-Hastings, K., Halter, D., Person, N., & Tutoring Research Group. (2000). Using latent semantic analysis to evaluate the contributions of students in AutoTutor. *Interactive Learning Environments, 8*(2), 129–147.

Heller, J. I., & Reif, F. (1984). Prescribing effective human problem-solving processes: Problem descriptions in physics. *Cognition and Instruction, 1,* 177–216.

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *The Physics Teacher, 30,* 141–158.

Hewitt, P. G. (1987). *Conceptual physics.* Menlo Park, CA: Addison-Wesley.

Hu, X., Cai, Z., Louwerse, M. M., Olney, A., Penumatsa, P., Graesser, A. C., et al. (2003). A revised algorithm for latent semantic analysis. In G. Gottlob & T. Walsh (Eds.), *Proceedings of the 2003 International Joint Conference on Artificial Intelligence* (pp. 1489–1491). Menlo Park, CA: AAAI Press.

Jordan, P., Makatchev, M., Pappuswamy, U., VanLehn, K., & Albacete, P. (2006). A natural language tutorial dialogue system for physics. In *Proceedings of FLAIRS 2006.* Menlo Park, CA: AAAI Press.

Jordan, P., Makatchev, M., & VanLehn, K. (2004). Combining competing language understanding approaches in an intelligent tutoring system. In J. C. Lester, R. M. Vicari, & F. Praguacu (Eds.), *Intelligent Tutoring Systems: 7th International Conference, ITS 2004* (pp. 346–357). Berlin, Germany: Springer-Verlag.

Jordan, P., Rosé, C. P., & VanLehn, K. (2001). Tools for authoring tutorial dialogue knowledge. In J. D. Moore, C. Redfield, & W. L. Johnson (Eds.), *Artificial intelligence in education: AI-Ed in the wired and wireless future* (pp. 222–233). Washington, DC: IOS.

Jordan, P., & VanLehn, K. (2002). Discourse processing for explanatory essyas in tutorial applications. In *Proceedings of the third SIGdial Workshop on Discourse and Dialogue* (pp. 74–83). Philadelphia, PA: Association for Computational Linguistics.

Katz, S., Connelly, J., & Allbritton, D. (2003). Going beyond the problem given: How human tutors use post-solution discussions to support transfer. *International Journal of Artificial Intelligence in Education, 13,* 79–116.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes, 25,* 259–284.

Lane, H. C., & VanLehn, K. (2005). Teaching the tacit knowledge of programming to novices with natural language tutoring. *Computer Science Education, 15*(3), 183–201.

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

Litman, D., Rosé, C., Forbes-Riley, K., VanLehn, K., Bhembe, D., & Silliman, S. (2006). Spoken versus typed human and computer dialogue. *International Journal of Artificial Intelligence and Education, 16,* 145–170.

Makatchev, M., Hall, B. S., Jordan, P., Papuswamy, U., & VanLehn, K. (2005, August). *Mixed language processing in the Why2-Atlas tutoring system.* Paper presented at the Proceedings of the Workshop on Mixed Language Explanations in Learning Environments, AIED2005, Amsterdam, the Netherlands.

Makatchev, M., Jordan, P., & VanLehn, K. (2004). Abductive theorem proving for analyzing student explanations and guiding feedback in intelligent tutoring systems. *Journal of Automated Reasoning, 32,* 187–226.

McNamara, D. S. (2001). Reading both high-coherence and low-coherence texts: Effects of text sequence and prior knowledge. *Canadian Journal of Experimental Psychology, 55,* 51–62.

McNamara, D. S., Kintsch, E., Songer, N., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from texts. *Cognition and Instruction, 14,* 1–43.

Mean, H. (1979). *Learning lessons: Social organization in the classroom.* Cambridge, MA: Harvard University Press.

Michael, J., Rovick, A., Glass, M. S., Zhou, Y., & Evens, M. (2003). Learning from a computer tutor with natural language capabilities. *Interactive Learning Environments, 11,* 233–262.

Olde, B. A., Franceschetti, D., Karnavat, A., Graesser, A. C., & Tutoring Research Group. (2002). The right stuff: Do you need to sanitize your corpus when using latent semantic analysis? In W. D. Gray & C. D. Schunn (Eds.), *Proceedings of the 24th Annual Conference of the Cognitive Science Society* (pp. 708–713). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Olney, A., Louwerse, M. M., Mathews, E. C., Marineau, J., Hite-Mitchell, H., & Graesser, A. C. (2003). Utterance classification in AutoTutor. In *Proceedings of the HLT–NAACL 03 Workshop on Building Educational Applica-*

tions Using Natural Language Processing (pp. 1–8). Philadelphia, PA: Association for Computational Linguistics.

Otero, J., & Graesser, A. C. (2001). PREG: Elements of a model of question asking. *Cognition and Instruction, 19,* 143–175.

Person, N., Graesser, A. C., Bautista, L., Mathews, E. C., & Tutoring Research Group. (2001). Evaluating student learning gains in two version of AutoTutor. In J. D. Moore, C. Redfield, & W. L. Johnson (Eds.), *Artificial intelligence in education: AI-ED in the wired and wireless future* (pp. 268–293). Amsterdam: IOS.

Person, N., Graesser, A. C., Kreuz, R. J., Pomeroy, V., & Tutoring Research Group. (2001). Simulating human tutor dialog moves in AutoTutor. *International Journal of Artificial Intelligence in Education, 12,* 23–39.

Person, N., Graesser, A. C., & Tutoring Research Group. (2002). Human or computer?: AutoTutor in a bystander Turing test. In S. A. Cerri, G. Gouarderes, & F. Paraguaca (Eds.), *Intelligent Tutoring Systems, ITS 2002* (pp. 821–830). Berlin, Germany: Springer.

Ploetzner, R., & VanLehn, K. (1997). The acquisition of informal physics knowledge during formal physics training. *Cognition and Instruction, 15,* 169–206.

Ranney, M., & Thagard, P. (1988). Explanatory coherence and belief revision in naive physics. In V. Patel & G. J. Groen (Eds.), *The Tenth Annual Conference of the Cognitive Science Society* (pp. 426–432). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Reif, F., & Scott, L. A. (1999). Teaching scientific thinking skills: Students and computers coaching each other. *American Journal of Physics, 67,* 819–831.

Reiner, M., Slotta, J. D., Chi, M. T. H., & Resnick, L. B. (2000). Naive physics reasoning: A commitment to substance-based conceptions. *Cognition and Instruction, 18,* 1–34.

Renkl, A. (2002). Worked-out examples: Instructional explanations support learning by self-explanations. *Learning and Instruction, 12,* 529–556.

Renkl, A., & Atkinson, R. K. (2003). Structuring the transition from example study to problem solving in cognitive skill acquisition: A cognitive load perspective. *Educational Psychologist, 38,* 15–22.

Rosé, C. P., Bhembe, D., Siler, S., Srivastava, R., & VanLehn, K. (2003). Exploring the effectiveness of knowledge construction dialogues. In U. Hoppe, F. Verdejo, & J. Kay (Eds.), *Artificial intelligence in education: Shaping the future of learning through intelligent technologies* (pp. 497–499). Amsterdam, the Netherlands: IOS Press.

Rosé, C. P., Jordan, P. W., Ringenberg, M., Siler, S., VanLehn, K., & Weinstein, A. (2001). Interactive conceptual tutoring in Atlas-Andes. In J. D. Moore, C. Redfield, & W. L. Johnson (Eds.), *Artificial intelligence in education: AI-Ed in the wired and wireless future* (pp. 256–266). Washington, DC: IOS.

Rosé, C. P., Moore, J. D., VanLehn, K., & Allbritton, D. (2001). A comparative evaluation of Socratic versus didactic tutoring. In J. D. Moore & K. Stenning (Eds.), *Proceedings of the Twenty-Third Annual Conference of the Cognitive Science Society* (pp. 897–902). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Rosé, C. P., Roque, A., Bhembe, D., & VanLehn, K. (2002). A hybrid language understanding approach for robust selection of tutoring goals. In S. A. Cerri, G. Gouarderes, & F. Paraguacu (Eds.), *Intelligent Tutoring Systems, 2002: 6th International Conference* (pp. 552–561). Berlin, Germany: Springer.

Sinclair, J. M., & Coulthard, R. M. (1975). *Towards an analysis of discourse: The English used by teachers and pupils.* London: Oxford University Press.

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

Slotta, J. D., Chi, M. T. H., & Joram, E. (1995). Assessing students' misclassifications of physics concepts: An ontological basis for conceptual change. *Cognition and Instruction, 13,* 373–400.

Swanson, J. H. (1992, April). *What does it take to adapt instruction to the individual? A case study of one-to-one tutoring.* Paper presented at the American Education Research Association, San Francisco, CA.

VanLehn, K. (in press). When is tutoring more effective than less interactive instruction? In M. T. H. Chi (Ed.), *Festshrift for Lauren Resnick.*

VanLehn, K., Jordan, P., Rosé, C. P., Bhembe, D., Bottner, M., Gaydos, A., et al. (2002). The architecture of Why2-Atlas: A coach for qualitative physics essay writing. In S. A. Cerri, G. Gouarderes, & F. Paraguacu (Eds.), *Intelligent Tutoring Systems, 2002, 6th International Conference* (pp. 158–167). Berlin, Germany: Springer.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes.* Cambridge, MA: Harvard University Pres.

Wood, D. J., & Middleton, D. (1975). A study of assisted problem-solving. *British Journal of Psychology, 66,* 181–191.

Wood, D. J., Wood, H., & Middleton, D. (1978). An experimental evaluation of four face-to-face teaching strategies. *International Journal of Behavioral Development, 1,* 131–147.

Appendix A. Problems

Training problems

1. Suppose you are running in a straight line at constant speed. You throw a pumpkin straight up. Where will it land? Explain.

2. The sun pulls on the earth with the force of gravity and causes the earth to move in orbit around the sun. Does the earth pull equally on the sun? Defend your answer.

3. Suppose you are in free falling elevator and you hold your keys motionless in front of your face and then let go. What will be the position of keys relative to your face as the time passes? Explain.

4. When a car without headrests on the seats is struck from behind, the passengers often suffer neck injuries. Why?

5. A clown is riding a unicycle in a straight line. She accidentally drops an egg beside her as she continues to move with constant velocity. Where will the egg land relative to the point where the unicycle touches the ground? Explain.

6. If a lightweight car and a massive truck have a head-on collision, upon which vehicle is the impact force greater? Which vehicle undergoes the greater change in its motion? Defend your answers.

7. A plane is supposed to drop leaflets on a crowd gathered to watch a swimming competition. The pilot times the drop so that the stack of leaflets is directly above the center of the pool when it is released. Unfortunately, someone forgot to take the stack out of the package and the pilot ends up dropping the sealed heavy packet of leaflets. Does the packet hit the center of the swimming pool? Explain.

8. Two closed containers look the same, but one is packed with lead and the other with a few feathers. How could you determine which had more mass if you and the containers were orbiting in a weightless condition in outer space?

9. Suppose that you released 3 identical balls of clay in a vacuum at exactly the same instant. They would all hit the ground at the same instant. Now you stick two of the balls together, forming one ball that is twice as heavy as the remaining, untouched clay ball. Both balls are released in a vacuum at exactly the same instant. Which ball hits the ground first?

10. If a car is able to accelerate at 2 m/s^2, what acceleration can it attain if it is towing another car of equal mass?

Essay Test A problems

A cat walks to the roof-edge of a building and just drops off. She takes 20 seconds to land on the ground. How does her velocity after the first half period of the fall compare with that on landing (air resistance is negligible)? Explain.

You observe two rocket ships in deep space where they are weightless. They have the same motors, which are generating exactly the same thrust force, but one ship is speeding up quickly while the other is speeding up slowly. Explain how this can be true.

A layer of water on the roof of a high building is frozen so as to provide a smooth icy horizontal surface. Two small closed boxes, one empty and the other filled with sand, are pushed such that they have equal velocity at the instant of falling off the edge. A little later, they land on the flat horizontal ground below. How will the distances of their landing points from the foot of the building be related? Explain, stating the principles of physics involved.

A diver jumps off a high platform. During the fall, he is being pulled down by the force of earth's gravity. Is there a force on earth due to the diver? If so, what is the earth's acceleration? Explain, stating the principles of physics involved.

Essay Test B problems

On the surface of the moon, a steel ball and a cotton ball are dropped at the same time from the same height. What would be the relation between the velocity they respectively acquire on reaching the moon's surface? Explain.

The Olympic rocket sled competition of 3002 is held in deep space, where gravity is negligible and there is no air resistance. The qualifying race is simply to start at rest, race exactly 100 km in a straight line, turn around, and return. One of the sled riders, Barry, gets sick at the last moment. He needs to choose a replacement rider from among his two friends, Maurice and Little Joe. Maurice is much larger than Little Joe, and on earth Maurice would weight more. However, they both fit inside the sled easily and they are equally skilled sled riders. Does it matter which rider Barry chooses? Explain.

The driver of a speeding truck finds that the truck brakes have failed just as he approaches the edge of a cliff. Rather than fly off the cliff in his truck, he opens the truck door and jumps out horizontally and perpendicular to the direction of motion of the truck just as the truck reaches the cliff-edge. Is he expected to land on the cliff? Explain.

A hiker claims that she can get out of any difficult situation! As a challenge, she is picked up by a helicopter and put in the middle of a frozen, icy pond and asked to reach the edge of the pond. The ice is so smooth and frictionless that when she tries to walk, her feet slide on the ice but her body stays where it is. She does some quick thinking, and then throws her helmet away, horizontally, as hard as she can. Will this help her get to the shore? Explain.

Appendix B. The Canned Text Remediation's Minilessons for the Truck–Car Problem

Here is an important point that a good explanation should cover. When the truck and car collide, they each exert a force on the other. This pair of forces is called an action/reaction force pair. Let us consider some general properties of reaction force pairs. From Newton's Third Law we know every action force causes an equal and opposite reaction force.

Furthermore, we know that the truck exerts a force on the car because it is in contact with the car, and in return the car exerts a reaction force because it is in contact with the truck. Thus,

since both the action force and the reaction force occur as a result of contact between the two bodies, both forces are contact forces. Thus, they have the same type. In general, an action force and its reaction force have the following properties: they are the same type, they are opposite to each other indirection, and they are of equal magnitude. Therefore, when the car exerts a force on the truck, at the same time the truck exerts an equal and opposite reaction force on the car.

Here is an important point that a good explanation should cover. The name of the law or principle of physics that you can apply to determine the difference in acceleration between the car and truck given that they both experience the same force is Newton's Second Law. We know that Newton's Second Law expresses a relationship between force, mass, and acceleration. The equation describing this relationship is Force equals Mass times Acceleration. Thus, Newton's Second Law states that force equals mass times acceleration. Applying this principle, if you apply equal force to two objects with different masses, you can use the relative accelerations of the two objects to determine the relationship between their respective masses. Therefore in this case, the vehicle with larger mass will have a smaller acceleration. Therefore you now can see that during the impact, the magnitude of the acceleration of the truck is less than the magnitude of the acceleration of the car at every time instance.

Here is an important point that a good explanation should cover. Just to be sure this point is absolutely clear to you, whenever you have an action/reaction force pair, the magnitude of the action force and the magnitude of the reaction force are equal. As an illustration, let's think about the forces involved when you hit a brick wall with your fist. Your fist exerts a force on the wall, and the wall exerts a reaction force on your hand. If you were to hit the wall harder, you can expect that it will hurt more. Thus the magnitude of the action force due to your hand is equal to the magnitude of the reaction force due to the wall. In general therefore the magnitude of the action force and the magnitude of the reaction force are equal.

Here is another relevant point. Let's think about the force exerted by your fist when it hits a brick wall. This scenario may help you remember the relationship between an action/reaction force pair. When you hit the wall, the direction of the force exerted on the wall by your fist is towards the wall. When your fist hits the wall it hurts. The force are you feeling when your fist hits the wall is the force exerted by the wall on your fist. The direction of the force exerted by the wall is towards your fist. So when your fist exerts a force on the wall, the wall exerts a force on your fist. The force towards your fist exerted by the wall is called the reaction force of the action force exerted by your fist on the wall. Together these two forces are known as a reaction force pair.

Now let's consider some general properties of reaction force pairs. The direction of a force and the direction of its reaction force are in opposition. Secondly, the force of your fist acting on the wall is a contact force. Contact forces, like that exerted by your fist on the wall, result from the physical contact between two objects. Therefore the force exerted by the wall on your fist is a contact force as well.

As for the magnitude of the action force and the magnitude of the reaction force, even though we cannot correctly deduce the correct answer from observation, we may be able to guess it. Notice that if you hit the wall harder, your fist will hurt more. Given this, what you can guess is that the magnitude of a force and the magnitude of its reaction force are equal. Therefore, in general a force and its reaction force have the same type, equal magnitude, and opposite direction.

Appendix C. Multiple choice questions from Test A addressing Newton's third law (Correct answer marked with *)

As a truck moves along the highway at constant speed, a nut falls from a tree and smashes into the truck's windshield. If the truck exerts a 1,000 N force on the nut, what is the magnitude of the force that the nut exerts on the truck?

- a) 1,000 N *
- b) less than 1,000 N
- c) 0 N (the nut does not exert a force on the truck)
- d) greater than 1,000 N (because the nut hit the truck, it exerts a greater force on the truck than the truck exerts on the nut)

An ocean liner traveling due east collides with a much smaller yacht, traveling due west. During the collision, the front end of the yacht is smashed in (causing the yacht to sink and the passengers to evacuate to their lifeboat). The ocean liner merely suffered a dent. Which is true of the relationship between the force of the ocean liner on the yacht and the force of the yacht on the ocean liner?

- a) because the yacht's acceleration during the collision was greater than the ocean liner's acceleration, the force of the yacht on the ocean liner is greater than the force of the ocean liner on the yacht
- b) the force of the ocean liner on the yacht is greater than the force of the yacht on the ocean liner
- c) the force of the ocean liner on the yacht is equal to the force of the yacht on the ocean liner *

You are in a seat in a rollercoaster when it accelerates forward, causing you to be pressed against the back of your seat. While the rollercoaster accelerates forward and you are pressed against the back of your chair, which of the following is true:

- a) there is a force on you in the forward direction *
- b) there is a force on you in the backward direction (opposite the direction you are moving in)
- c) there are no forces acting on you

A 30-kg child receives her first "A+" on a spelling test and, overcome with joy, jumps up and down her 200-kg desk. This desk is very strong and does not move while the child jumps on it. Which of the following is true?

- a) the child exerts a force on the desk but because it does not move, the desk does not exert a force on the child
- b) the desk exerts a force on the child but the child does not exert a force on the desk
- c) the child and desk both exert a force on the other *

A distant planet orbits a nearby star. Which of the following statements is true:

- a) The star exerts a gravitational force on the planet, but the planet does not exert a gravitational force on the star

b) Both the star and the planet exert a gravitational force on the other, but the force of the star on the planet is greater than the force of the planet on the star
c) Both the star and the planet exert a gravitational force on the other, and the gravitational force of the planet on the star is the same as the force of the star on the planet *

Two billiard balls of equal size and weight initially traveling in opposite directions collide with each other. The magnitude of the force the first ball exerts on the second is 3 N. What is the magnitude of the force of the second ball on the first?

a) 3 N *
b) Less than 3 N
c) More information is necessary to answer this question

If the direction of the force of the first ball on the second ball is to the right, what is the direction of the force of the second ball on the first?

a) Also to the right
b) To the left *
c) More information is necessary to answer this question