REFLECTION AND LEARNING ROBUSTNESS IN A NATURAL LANGUAGE CONCEPTUAL PHYSICS TUTORING SYSTEM

by

Arthur Ward

MBA, Carnegie Mellon University

M.S., Intelligent Systems, University of Pittsburgh

Submitted to the Graduate Faculty of

the Intelligent Systems Program in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2010

UNIVERSITY OF PITTSBURGH

INTELLIGENT SYSTEMS PROGRAM

This dissertation was presented

by

Arthur Ward

It was defended on

June 23rd 2010

and approved by

Diane Litman, Intelligent Systems Program, University of Pittsburgh

Sandra Katz, Learning Research and Development Center, University of Pittsburgh

Alan Lesgold, Intelligent Systems Program, University of Pittsburgh

Christian Schunn, Intelligent Systems Program, University of Pittsburgh

Dissertation Director: Diane Litman, Intelligent Systems Program, University of Pittsburgh

REFLECTION AND LEARNING ROBUSTNESS IN A NATURAL LANGUAGE CONCEPTUAL PHYSICS TUTORING SYSTEM

Arthur Ward, PhD

University of Pittsburgh, 2010

This thesis investigates whether reflection after tutoring with the Itspoke qualitative physics tutoring system can improve both near and far transfer learning and retention. This question is formalized in three major hypotheses. H1: that reading a post-tutoring reflective text will improve learning compared to reading a non-reflective text. H2: that a more *cohesive* reflective text will produce higher learning gains for most students. And H3: that students with high domain knowledge will learn more from a less cohesive text.

In addition, this thesis addresses the question of which *mechanisms* affect learning from a reflective text. Secondary hypotheses H4 and H5 posit that textual cohesion and student motivation, respectively, each affect learning by influencing the amount of inference performed while reading.

These hypotheses were tested by asking students to read a reflective/abstractive text after tutoring with the Itspoke tutor. This text compared dialog parts in which similar physics principles had been applied to different situations. Students were randomly assigned among two experimental conditions which got "high" or "low" cohesion versions of this text, or a control condition which read non-reflective physics material after tutoring. The secondary hypotheses were tested using two measures of cognitive load while reading: reading speeds and a self-report measure of reading difficulty.

Near and far transfer learning was measured using sets of questions that were mostly isomorphic vs. non-isomorphic to the tutored problems, and retention was measured by administering both an immediate and a delayed post-test. Motivation was measured using

a questionnaire.

Reading a reflective text improved learning, but only for students with a middle amount of motivation, confirming H1 for that group. These students also learned more from a more cohesive reflective text, supporting H2. Cohesion also affected high and low knowledge students significantly differently, supporting H3, except that high knowledge students learned best from high, not low cohesion text.

Students with higher amounts of motivation did have higher cognitive load, confirming hypothesis H5 and suggesting that they engaged the text more actively. However, secondary hypothesis H4 failed to show a role for cognitive load in explaining the learning interaction between knowledge and cohesion demonstrated in H3.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1.0 INTRODUCTION

Physics education, as well as science education generally, is faltering in the United States. Although fourth grade American students score slightly above average for OECD (Organization for Economic Co-operation and Development) countries in science literacy, by the time they are 15 their scores have fallen into the bottom third (Provasnik et al., 2009) of all OECD countries.

Although improving these statistics has long been recognized as a national educational priority, progress has been difficult. Between 1995 and 2007, there was no measurable improvement in American science scores (Provasnik et al., 2009). There is still a great need for educational interventions that can improve learning outcomes in math and science.

It is broadly recognized that one of the most effective educational interventions is one-on-one tutoring with a human tutor (Bloom, 1984; Kulik et al., 1990). Individualized tutoring has been reported to produce large learning gains, in the range of two full standard deviations (Bloom, 1984). The limited supply and high cost of talented human tutors, however, has made necessary another educational goal: the creation of effective Intelligent Tutoring Systems (ITSs). In a speech to the National Academy of Sciences in 2009 (Obama, 2009), President Obama envisioned "learning software as effective as a personal tutor."

Creating software tutors that fulfill this vision is still an un-obtained research goal, but for the reasons noted above, an important focus of this research is the area of science literacy.

One difficult area of science education is *qualitative* physics. Qualitative physics emphasizes the conceptual understanding of physics principles, rather than quantitative problem solving. Research in qualitative physics is motivated by the observation that students can do well in physics class, and become competent solvers of quantitative physics problems, but still have a poor conceptual understanding of physics and retain many misconceptions (Hal-

loun and Hestenes, 1985b,a). An important consequence of poor conceptual understanding is a reduced ability to transfer learning to new problem solving situations. "Transfer" is the use of learned material in situations other than that in which it was learned.

The issue of poor conceptual physics understanding has been investigated using several tutoring systems which specialize in qualitative physics, for example Why2-Atlas (VanLehn et al., 2002), Why2-Autotutor, (Graesser et al., 2005) and Itspoke (Litman and Silliman, 2004). This study uses the Itspoke system (Litman and Silliman, 2004), which will be described more fully in Section 3.1. Briefly, Itspoke is a speech enabled version of the Why2-Atlas (VanLehn et al., 2002) text based tutor. It teaches by first presenting students with an introductory text about physics. Then, after a pre-test, it engages them in a series of spoken tutoring dialogs about each of five problems in qualitative physics, followed by a post-test.

The issue of transfer has been studied separately by other researchers (Gick and Holyoak, 1983), who have found that a process of *abstractive* comparison of previous problem solving episodes can help in solving dissimilar target problems. The success of comparative abstraction in improving transfer in other fields raises the question of whether it could also improve conceptual understanding and transfer in the domain of qualitative physics.

In this study I implement a form of comparative abstraction by administering a reflective reading to students following their tutoring sessions with the Itspoke tutor. This reading points out and compares parts of the previous problem solving episodes in which similar physics laws were applied. The hypothesis is that, similar to results in other domains (Gick and Holyoak, 1983), this comparison will help students infer what parts of the previously tutored problems were fundamental and important, and which were incidental surface features. This should result in improved learning and transfer.

Other research (e.g. (Katz et al., 2003, 2007)) suggests that reflection can be an effective way to improve learning after tutoring, and also that reading a reflective text, such as the one used here, will be sufficient to produce learning gains.

The decision to use a reflective text, however, raises the issue of how that text should be structured to maximize learning gains. One important and frequently studied feature of text is its *cohesion*. Cohesion is the property of a text that makes it seem to "hang

together." [1] Textual cohesion has been shown to significantly affect how well the text is understood by the reader. For most readers (Britton and Gulgoz, 1991; McKeown et al., 1992), higher cohesion seems to make a text easier to comprehend. However McNamara and colleagues (McNamara et al., 1996; McNamara and Kintsch, 1996; McNamara, 2001; O'Reilly and McNamara, 2007) have shown that for certain students, low cohesion text can actually produce higher comprehension. They hypothesized that this is because low cohesion text includes gaps which the reader must use inference to bridge, and these inferences cause learning. This explanation implies a higher cognitive load for these students when reading lower cohesion text.

Another factor that has been shown to influence text comprehension is reader interest (Schiefele, 1996). Higher levels of interest seem to be associated with better comprehension, also possibly because of more active processing. This explanation for the benefits of interest also implies higher cognitive load for more interested students when reading.

These considerations lead to three primary hypotheses which are tested in this work. H1 hypothesizes that administering an abstractive/reflective reading after tutoring with the Itspoke tutor will improve learning and transfer in qualitative physics. H2 suggests that the cohesiveness of this reading will impact its superiority over a non-reflective control. For most students, high cohesion should cause more learning. H3 holds that the effect of textual cohesion will vary with knowledge level. Following McNamara, I expect high pre-testers to learn more from a low cohesion text.

In addition, the hypothesized mechanisms underlying the effects of textual cohesion and interest suggest two secondary hypotheses. H4, that textual cohesion affects learning by modifying the amount of inference made from text. H5, that motivation also affects inference from text. Both of these hypotheses imply differences in cognitive load. If low cohesion causes more inference, it should be accompanied by increases in cognitive load. Similarly, if higher motivation causes more inference while reading text, it should also be accompanied by higher cognitive load.

These hypotheses were tested by asking students to read a text following tutoring with the Itspoke tutor. In the control condition students read a shortened version of the non-reflective

[1] More formal definitions of cohesion will be discussed in Sections 2.4 and 6.

introductory physics text. In the experimental conditions, students read either a high or low cohesion version of a reflective text, which compared relevant parts of the preceding tutorial dialogs. Learning gains were measured using pre-tests, post-tests and delayed post-tests. The readings were all administered in a tap-to-read interface which allowed the collection of both self-report and reading speed measures of cognitive load. In addition, student motivation was assessed using a questionnaire patterned on work by Pintrich and DeGroot (1990).

Results showed that this type of abstractive/reflective text did significantly improve learning after qualitative physics tutoring, as measured by both immediate and delayed post-tests, but only for students with a middle level of motivation, supporting H1 in this group. Students in the highest or lowest motivation groups did not benefit from reflection. In general, reflective text with higher cohesion caused higher learning gains than low cohesion text, in this group, supporting H2. However, results also showed a significant interaction between knowledge level and textual cohesion. This supports H3, except that middle motivation students with low domain knowledge had higher mean learning gain from low cohesion text, and students with high domain knowledge had higher mean learning gain from high cohesion text.

For the secondary hypotheses, results using the cognitive load measures suggest that students with higher motivation did engage the text more actively than those with lower motivation, supporting H5. However H4 was not supported: there were no significant cognitive load differences to explain the interaction found for middle motivation students in H3.

2.0 BACKGROUND AND MOTIVATION

In this thesis I describe an experiment using the Itspoke tutor, which teaches *qualitative* physics. Itspoke will be described in more detail in Section 3.1. In this section I begin by describing the general concern with problem solving based tutors that was mentioned in the introduction: they don't promote robust conceptual learning, which leads to poor transfer. I next briefly describe various tutors that have been designed to address this problem, including Itspoke, the one used in this study. Then I describe a body of research into the transfer of learning, which suggests that producing an abstract representation is important for achieving robust learning and transfer. In Section 2.3 I survey work in reflection and present a taxonomy of the various types that have been studied. I argue that this work suggests that a carefully designed reflective reading could aid student abstraction and so help transfer in conceptual physics tutoring. This work informs decisions about how to augment Itspoke with a reflective reading, as well as suggesting the structure and the content of the reading.

In Sections 2.4, 2.5 and 2.6, I review work suggesting a set of important factors that should be considered when evaluating the effectiveness of a reflective text. These factors are the cohesiveness of the text, the effect of textual cohesion on inference, and student motivation. Finally, in Section 2.7 I describe specific hypotheses generated by these considerations.

2.1 QUALITATIVE PHYSICS TUTORING

Teachers in elementary mechanics courses have long noticed that students can do well in physics class, learn to competently solve quantitative physics problems, and still have a very

poor understanding of physics concepts. In several studies, students who had performed well in quantitative problem solving showed poor transfer to qualitative physics problems (Halloun and Hestenes, 1985b,a), indicating that quantitative performance had masked a shallow understanding of physics concepts. "Quantitative" physics emphasizes solving numerical problems, whereas "qualitative" physics emphasizes understanding how and why physics concepts are applied. Qualitative physics problems usually involve comparing magnitudes, rather than solving for numerical answers. Shallow learning with poor conceptual understanding diminishes a student's ability to properly use domain language, to "reality check" answers to physics problems, and to transfer problem solving skill to new problems. The problem of shallow learning in physics is not limited only to classroom instruction.

VanLehn et al. (2000) noted that model tracing tutors have also been criticized for failing to encourage deep learning. One criticism is that tutors don't "promote stepping back to see the 'basic approach' one has used to solve a problem." In addition, they note that quantitative problem solving skills do not transfer well to qualitative problems. Students who do well in quantitative problem solving often also do poorly on measures of conceptual understanding such as the Force Concepts Inventory (Hestenes et al., 1992).

The problem of shallow learning described above has also been described as a lack of "robust learning." Robust learning is broadly defined to include both transfer and long term retention (PSLC, 2009). Retention refers to the persistence of learning, as measured, for example, by a delayed post-test. Transfer is understood to be what happens when knowledge is applied in a situation which is different than the one in which it had been learned. Given that we would like learning to be useful outside an academic setting, retention and transfer are some of the most important goals in tutoring research.

VanLehn et al. (2000) suggested that model tracing tutors should be improved to deal with the problem of shallow learning. In particular they suggested engaging students in a natural language dialog designed to help them infer or construct a deeper understanding of the target material. These interactive directed lines of reasoning were called "knowledge construction dialogs." They investigated the addition of knowledge construction dialogs to the Andes physics homework helper. Similar considerations led to a number of tutors which used natural language to encourage deeper conceptual understanding. These tutors included

Atlas-Andes (Rosé et al., 2001), Why2-Atlas (VanLehn et al., 2002), AutoTutor (Graesser et al., 2005), and Itspoke (Litman and Silliman, 2004). Itspoke is described in more detail in Section 3.1. By emphasizing conceptual learning, these tutors sought to produce a deeper and more transferable understanding of physics. All of these tutors have produced learning gains. For example, Rosé et al (2001) found that a version of the Andes model tracing tutor which included short dialogs about conceptual physics improved learning of physics concepts compared to a problem-solving only version of the tutor, with an effect size of 0.9 standard deviations.

These learning gains do not seem as strong as those reported for human tutors (Bloom, 1984), however, so there may still be room for improvement. Hints about an additional source of improvement come from the transfer literature, which has investigated a method of producing transfer in non-physics domains which may be complementary to the approaches described above.

2.2 ABSTRACTION AND TRANSFER

As described above, qualitative physics tutors have had success in improving learning gains by using natural language dialog to teach abstract physics concepts directly. As I will describe in this section, work in other domains has shown that transfer can be increased by helping students induce more abstract representations by comparing examples. This section provides an overview of that work, and argues that a similar process of comparison and abstraction might increase learning and transfer in qualitative physics.

The role of abstraction has been widely studied in the transfer literature, notably by Gick and Holyoak (Gick and Holyoak, 1983). They considered "Dunker's radiation problem," which is a standard problem in transfer research. The students are asked to find a way to irradiate a tumor with enough radiation to kill it, without harming surrounding tissue (the answer is to divide the radiation into several beams which converge on the tumor). Before solving this target problem, students are exposed to an analogous problem, such as a general dividing an army to attack a fortress from several directions. Typically students can learn,

understand and remember the source analog, but still show very poor transfer to the target problem. Gick and Holyoak asked students to summarize the source analog, or to state its underlying principle, or to make a diagram of the problem, all without significantly improving transfer. However, when they gave students two analogs and asked them to describe their similarities, transfer was (finally) improved. Their interpretation for these results was that students were using a process of induction to create more general, abstract representations of the problem, which were more easily applied to the target problem.

Having a more general and abstract representation of knowledge is generally thought to be important for transfer. These representations are often thought of as mental schemata (Reed, 1993), which can contain both some specific contextual details of the learning context, and also more general information. Schemata are formed by a process of induction from the learned material and other sources.

A number of instructional factors are thought to affect schema induction from examples (Gick and Holyoak, 1987). These include the number of examples and the order in which they are presented, as well as the presence of abstract training. The process of schema induction is also mediated by domain knowledge (Chi et al., 1982), with experts structuring their knowledge according to deeper domain principles, and novices creating structures more dependant on surface features.

Other individual differences such as I.Q. and memory span (Skanes et al., 1974; Leher and Littlefield, 1993; Goska and Ackerman, 1996) have also been found to affect learners' ability to create and use transferable schemata. Individual differences in background knowledge have also been shown to affect reader's ability to make inferences from text, as described in Section 2.4.

The work described above was all in non-physics domains, however my own previous work in tutorial dialog cohesion has suggested that abstraction is also an important mechanism in learning from physics tutoring. In a series of studies, I measured cohesion in tutorial (non reflective) dialog automatically by counting the number of "cohesive ties" (Halliday and Hasan, 1976) between tutor and student. A cohesive tie was counted whenever the tutor or student repeated each other's choice of a word or word-stem (Ward and Litman, 2006). In later work (Ward and Litman, 2008), the measure was extended to count cohesive ties

between words which had semantic relationships to each other as measured by WordNet's (Miller et al., 1990) hyponym/hypernym hierarchy. In (Ward and Litman, 2008) we found that cohesive ties that repeated the other participant's usage but at a higher or lower level of abstraction improved correlations with learning, particularly with far transfer learning.

The results described above were obtained by automatically recognizing the semantic relationships between words. We next investigated if the relationship between tutor-student cohesion and learning would also be present with (presumably) more accurate human tagging of the cohesive ties. In (Ward et al., 2009), we tagged a corpus of reflective tutorial dialogs for various types of cohesive tie. We again found that abstractive ties, in which one participant repeated the other's contribution but at a greater level of generality, were correlated with learning. We also found that the reverse type of tie, in which repetition was at a more specific level, was correlated with learning.

We have seen that inducing abstract schema from examples can help learning and transfer in non-physics domains, and also that abstraction is an important mechanism for learning from tutoring dialogs with our physics tutor. This suggests that abstracting from multiple examples may also be an effective intervention in physics tutoring, however there are reasons to think it may be difficult to do during problem solving. For example, working memory constraints may make it difficult to induce while simultaneously remembering the state of the current solution. In addition, practical considerations suggest that schema induction from several examples would be best attempted reflectively, after the examples have been presented.

2.3 REFLECTION AND ABSTRACTION

The previous section argued that the process of inducing abstract schema from examples, which has increased transfer in other domains, could also increase transfer in qualitative physics tutoring. It also suggested that this might be best done as a *reflective* process, after tutoring is complete. Pursuing that suggestion, in this section I examine some related work which provides information about how best to implement reflection. This work provides

a useful taxonomy of reflection, which allows us to categorize the type of post-tutoring reflection suggested here in relation to other work. It also suggests that reading a reflective text may be as effective as more interactive forms of reflection, and it provides a theoretical description of the reflective process which will be useful in determining the structure of the reflective text.

Tchetagni, Nkambou, and Bourdeau (2007) draw a distinction between "reflection-in-action" and "reflection-on-action [1]" According to these authors, reflection-in-action is when a student reflects on problem solving activity *during* that activity. Reflection-on-action is used to mean when a student reflects *after* the problem solving activity. By this scheme, the kind of abstractive reflection suggested in the previous section is "reflection-on-action."

Next, I briefly review work in the reflection-on-action category, which suggests that a reflective reading after problem solving could help students create more abstract and transferable representations. Following that I describe some work in the reflection-in-action category, which suggests a useful way to structure the reading.

2.3.1 Reflection-on-action

Reflection-on-action has a large scope (Tchetagni et al., 2007), which considers an entire solution path or the student's cognitive state after problem solving. Collins and Brown (1986) provide a very useful way to subdivide the category of reflection-on-action. They point out that the computer can promote reflection by making the student's own thought processes and learning visible. They suggest several ways this might happen, which are paraphrased below:

1. Students can compare their own solution process to that of an expert
2. Students can see different portions of a process together, or see aspects of a process otherwise invisible
3. Students can derive abstractions about the process by comparing multiple performances simultaneously

[1] They take these terms from Donald Schön (1983).

4. Abstractions can be developed in a form that is helpful for developing good metacognitive strategies

The Sherlock II tutor (Katz et al., 1998) implemented reflection of the third type: the comparison of multiple performances simultaneously. Among other choices, students were able to select a replay of their own problem solving sequence, a replay of an expert's solution, or a comparison of their own and an expert's solution.

In Katz et al. (2003), study 1, tutors and students were directed to "reflect-on-action" after problem-solving sessions with the Andes tutor. The goal of this laboratory study was to describe student-tutor interaction during reflective dialogs about physics. Among other things, the human tutors would refer back to the previous problem solving sessions and offer generalizations of physics concepts and problem solving strategy. Interestingly, Katz et al. measured the amount of abstraction in these dialogs by counting the incidence of certain dialog events, including conceptual generalization, conceptual specialization and strategic generalization. They found that the amount of abstraction in them correlated with quantitative learning. They considered this a measure of near transfer, because quantitative learning was measured using problems similar to those that had been tutored in Andes.

In a second experiment, Katz et al. (2003) compared two types of reflection-on-action. In one condition they asked students reflective questions which were followed by interactive followup discussions with a human tutor. In a second condition, the questions were followed by a standard "canned text" response. They found that the canned text reflective response was just as effective as the interactive personalized response. This study is discussed in more detail in Section 7.

Similarly, in (Katz et al., 2007), Katz and colleagues also presented reflection questions after quantitative problem solving with the Andes physics tutor, but this time in a classroom rather than a laboratory based study. Student answers to these questions were followed by either an interactive dialog or a "canned" text response. Although reflection was shown to improve conceptual understanding of physics, there was again no advantage to the interactive dialog relative to the fixed text. In fact, the text response performed marginally better than dialog.

Reflection about the problem solving process has also been shown effective in the context of classroom active-learning environments. Davis and Linn (2000) compared giving students "self-monitoring" prompts, which asked them to plan or to reflect on their own learning vs. "activity" prompts, which encouraged them to complete task steps. They found that while activity prompts helped task completion, self-monitoring prompts led to greater knowledge integration.

In later work, Davis (2003) compared reflective prompts with various degrees of specificity. This work is described in greater detail in Section 7.

Together, these results confirm that reflection-on-action interventions have successfully increased learning, including in the domain of quantitative (but not qualitative) physics. In addition, the Katz studies surveyed imply that reflection need not be implemented using sophisticated personalized feedback. These results imply that reflection-on-action after tutoring with Itspoke, implemented using a text, could improve learning. This expectation is formalized as Hypothesis One in Section 2.7

In this work I add reflection of Collins and Brown's third type, above, to the Itspoke tutor. This reflection will be scaffolded by a reading which points out comparable places in different tutored solution paths, and discusses what aspects are common and which are unimportant surface details.

2.3.2 Reflection-in-action

The second type of reflection in Tchetagni et al's taxonomy is "reflection-in-action." This is done during problem solving, rather than after, and so is different from the kind implemented in this thesis. However, an implementation of reflection-in-action by Tchetagni et al. (2007) offers an example of a general four step reflective dialog frame, which I use to structure the current work's reflective reading.

Tchetagni et al. (2007) modify the remedial sub-dialogs used by their Prolog-Tutor to explicitly follow a four-step reflective dialog frame. The four steps are taken from philosophical work by John Dewey (1910), and could also apply to on-action reflection:

1. Elicit curiosity

2. Identify relevant facts of the learning situation
3. Draw a solution by linking facts to principles, concepts, heuristics or past experiences in the learning domain
4. Evaluate the correctness of the solution

Several aspects of this four-step reflective dialog frame are supported by more recent work in learning. For example the first step, "elicit curiosity" could provide a benefit similar to that of impasses, which have been shown to be important in tutoring (VanLehn et al., 2003). The second step reminds the student of relevant context which may have been forgotten at the end of problem solving. A contextualization step has also been used in a successful after-action-review tutor (Pon-Barry et al., 2005). The third step allows for showing how common principles apply to different sets of specific facts in the tutoring dialogs, which is similar to abstraction as discussed in Section 2.2. The fourth step involves task-oriented feedback, which has been shown to increase learning in several studies (Kluger and DeNisi, 1996).

In section 4.3.1, I describe how I used these steps to structure the reflective readings which were given after tutoring in the Itspoke tutor.

2.4 TEXTUAL COHESION AND LEARNING

As Section 2.2 argued, results from the learning literature suggest that causing students to induce abstract schema from several examples will be a good way to increase far-transfer learning. Section 2.3 further suggested that a good way to nurture this type of abstraction would be through a reflective reading.

In this section, I review work suggesting that certain a property of a reading, namely its *cohesion* can affect how well students are able to make inferences and build mental models from it. Based on these considerations, in Section 2.7 I hypothesize a role for cohesion in learning from a reflective text.

Cohesion is often defined (e.g.: (Morris and Hirst, 1991)) to be the degree to which text "hangs together." Halliday and Hasan (1976) propose that cohesion is generated by certain

lexical and syntactic features of the text such as word or synonym repetition, which are called "cohesive ties." Cohesive ties tend to make relationships in the text, such as logical, causal or temporal relationships, explicit. In the absence of cohesive ties, the relationships within a text have to be inferred by the reader. In the current work, as well as in previous papers (e.g.: (Ward and Litman, 2008)) I use the term "cohesion" to refer to these properties of the text, and "coherence" to refer to properties of the mental model constructed by the reader. This is a common usage in the text comprehension literature. For example, O'Reilly and McNamara apply a similar definition " the degree to which the concepts, relations and ideas within a text are explicit" to the term "cohesion" and define "coherence" as a property of the reader's comprehension. However there is also some variation in usage. In (McNamara, 2001) textual *coherence* is defined to be the "extent to which the relationships between ideas in a text are explicit." Also McKeown et al. (1992) use the term "coherence" in describing changes made to increase the cohesiveness of a text. In the following discussion, I will use the word "cohesion" throughout to make clear when I am referring to properties of a text, rather than of the mental model of the reader.

Several studies have shown that high cohesion text is, in general, more easily understood and remembered than low cohesion text. For example, Britton and Gulgoz (1991) increased textual cohesion by adding information where ever software based on Kintsch's reading comprehension model indicated inference was required. They found that students who read the revised version of the text remembered more, and were able to reproduce significantly more of the text's structure.

Similarly, McKeown et al. (1992) found that students who read a text revised for increased cohesion recalled significantly more, and did better on a post-test, than students who read a less cohesive version. Also McNamara et al. (1996) experiment 1, found that a more cohesive text improved recall overall.

More cohesive text has thus been shown to improve comprehension in general, but in a series of experiments (McNamara et al., 1996; McNamara and Kintsch, 1996; McNamara, 2001; O'Reilly and McNamara, 2007) McNamara and her colleagues have added additional detail to the picture. Specifically, they have shown that students with low domain knowledge react differently to textual cohesion than students with high domain knowledge.

In repeated experiments with several different texts, students with low domain knowledge have been shown to learn better from texts with *high* cohesion. Students with higher domain knowledge learned better from texts with *low* cohesion. This was interpreted to be because the low knowledge students are unable to make the inferences necessary to create a coherent mental representation of the low cohesion texts, and so only learn when given the additional cues from the high cohesion texts. The high knowledge students, on the other hand, get an "illusion of knowing" when reading the high cohesion text. They make few new inferences, and so don't learn. When reading the *low* cohesion text, the high knowledge students are triggered to make inferences by the gaps in the text. They are able to complete these inferences because of their higher knowledge level, and so they learn. McNamara's more recent work (O'Reilly and McNamara, 2007) suggests that the advantage of low cohesion text is specific to high knowledge readers with low comprehension skill. This is thought to be because readers with higher comprehension skill actively engage the text, and so draw inferences and learn even from the high cohesion version. Note however that the interaction of cohesion and prior knowledge does not appear in every study. For example, Boscolo and Mason (2003) failed to find such and interaction in their study of motivation and cohesion, which is described more fully in Section 2.6.

In previous work, I have shown a similar interaction between student domain knowledge and lexical cohesion, but in tutoring dialog rather than in text (Ward and Litman, 2006, 2008). In this case cohesion was measured as the number of cohesive ties between tutor and student. Cohesive ties were counted when tutor and student used the same words (or word stems) in adjacent turns. Like cohesion in text, dialog cohesion was found to be correlated with learning for the students with lower pre-test scores. This was also interpreted to be a function of domain knowledge: when students with lower domain knowledge used the same terms as their tutor, it was evidence that they had made new inferences about their meaning. When high knowledge students repeated their tutor's terms, however, it was not an indication of new inference, because they were already familiar with the domain. This work is described more completely in Section 6.

In many of McNamara's studies of the cohesion reversal effect, she has found that low cohesion text has different effects on different levels of textual representation, understood

in the sense of VanDijk and Kintsch's theory of text comprehension (vanDijk and Kintsch, 1983). In that theory, a reader builds several different levels of mental representation while reading a text. At the shallowest level, a verbatim representation of the words is built, and quickly decays. At the next level, a propositional model is built of those objects and relationships which are explicitly mentioned in the text. This representation decays more slowly than the verbatim representation. At the deepest level is the situation model. This representation is built from the propositional model using inference and world knowledge. This model is thought to be more elaborated and to last much longer in memory than the propositional model.

This situation model representation was shown to be improved by low cohesion text in several of McNamara's studies, which is relevant to our interest in robust learning[2]. As just described, the situation model is thought to be a richer representation of the text, which has been elaborated using real world knowledge. In our physics domain, this richer representation may be useful for solving "far transfer" problems, which are dissimilar to the tutored problems. Situation models are also thought to decay more slowly than shallower representations, and so may help with retention, the other aspect of robust learning.

As mentioned above, McNamara attributes the effect of low cohesion text to its influence on inference during reading. Other researchers, however (Kalyuga and Ayres, 2003), have suggested an alternative mechanism for the cohesion reversal effect. Kalyuga and Ayres (2003) suggest that readers with high domain knowledge fail to learn from a highly cohesive text because such a text conflicts with their pre-existing and relatively well developed domain schema. They must reconcile the text with their own schema, and this extraneous cognitive load inhibits learning. This explanation predicts the observed interaction between domain knowledge and textual cohesion, but also predicts a different level of cognitive load in the reader. If the "schema interference" explanation is correct, we should see a high cognitive load for high pre-testers reading highly cohesive text. McNamara's explanation is that the high pre-testers don't learn because they are not triggered to make any inferences from the

[2]Note, however in two studies (O'Reilly and McNamara, 2007; McNamara, 2001) the effects were instead seen in more shallow "text base" measures. This was interpreted to be because the materials used in those studies, which were about cell division, were so difficult that subjects could not form coherent situation models, and so fell back to text based representations.

high cohesion text. Under this explanation, cognitive load should be lower when reading the highly cohesive text than when reading the less cohesive text.

This difference is of importance for developers of tutoring systems, because the two explanations have different implications for how to treat a knowledgeable student. From a knowledge integration viewpoint (e.g.: (Davis and Linn, 2000; Davis, 2003)) schema reconciliation could be potentially beneficial. It may be possible to help make it more productive by reducing extraneous sources of cognitive load, or by removing parts of the instructional material which are simply redundant with the pre-existing schema. On the other hand, the "illusion of knowing" interpretation suggests that steps should be taken to further engage the student with the material, perhaps by presenting question-stem prompts (Bell and Davis, 2000) about the text.

This section has described evidence that the cohesiveness of the reflective text used in this study may impact how well students learn from it. The work of Britton (Britton and Gulgoz, 1991) and others suggests that higher cohesion may improve comprehension of the reflective text. However, McNamara's work suggests that low cohesion may be more beneficial for certain groups of students. These expectations are formalized as Hypotheses Two and Three in Section 2.7 on Page 21.

In an attempt to distinguish between the two potential explanations for the cohesion reversal effect, I add two simple measures of cognitive load to the tutoring environment. The expected effect of cohesion on cognitive load is described in Hypothesis Four, in Section 2.7. The measures of cognitive load are described next.

2.5 COGNITIVE LOAD

Cognitive load theory has been studied as a way to understand how the structure of educational material affects learning difficulty (Sweller, 1994). Typically, theorists divide cognitive load into several types. "Intrinsic load" is the load made necessary by the learning task, "extraneous load" is additional load imposed by instructional design, while "germane load" is additional load from the construction of mental schema (Paas et al., 2003). Accurately

measuring these various types of load is important to cognitive load theory.

Various measures of cognitive load have been tested in the literature (e.g. (Rubio et al., 2004)) including subjective measures which question the learner about load, and performance measures which measure aspects of a primary or secondary task. Subjective measures include likert scale questions about task difficulty, while performance measures include reading times.

Schultheis and Jameson (2004) compare reading time to several other measures of cognitive load while reading: subjective self report, P300 amplitude (a measure of the brains reaction to stimulus) and pupil size. A repeated measures Anova found that all the measures of cognitive load except pupil size were significantly different between easy and difficult texts.

Schultheis and Jameson (2004) used a 4 point scale to self-report cognitive load, in which 1 was labeled "easy" and 4 was labeled "difficult. " However, it seems to be common to use longer self-report scales. Paas et al. (2003) (Table 1) lists 27 cognitive load studies, of which 8 used a 7 point rating scale, and 16 used a 9 point scale (the other 3 used non self-report methods such as secondary task).

In the current work I use both subjective "self-report" and empirical "reading-time" measures of cognitive load to examine the relation between textual cohesion and inference. These measures were implemented in the Linger environment which will be described in Section 3.2. The measures are described in Section 4.4.2 and evaluated in Section 5.1.3.

2.6 STUDENT MOTIVATION

In this section I review evidence suggesting that student motivation is an important factor in learning. I first briefly review some related work in detecting and managing motivation during tutoring. Then, I describe an instrument used to measure motivation, and work linking motivation to learning both in the classroom and from text. Based on the results linking motivation to text comprehension, I then suggest measuring motivation in the current study, as a potential factor affecting the effectiveness of the reflective reading.

The majority of Intelligent Tutoring Systems which implement student models model only the student's cognitive state. That is, the system tries to keep track of what the

student knows, but not of how the student is feeling. In contrast to this approach, expert human tutors seem to actively manage the student's emotional state. For example, Graesser et al. (Graesser et al., 1995) state that a good human tutor "bolsters student motivation, confidence and self-efficacy while mastering the material."

Work has begun to also add this capability to computer tutors. For example, del Soldato (del Soldato and du Boulay, 1995) describes an architecture for tutoring systems which maintains both cognitive and motivational student models, and reconciles the difference when their recommendations disagree. Other researchers (e.g.(Litman and Forbes-Riley, 2006b, 2004; Dmello et al., 2005)) have done work suggesting that student emotional state can be detected during tutoring and so could be used to inform the student model. Also, Aist et al. (2003) have shown that a system using human-supplied emotional scaffolding helped students stay on task with an automated reading tutor for a longer period of time. This suggests that managing student motivation during tutoring could indeed have a payoff in learning gains.

In other studies, motivation is typically assessed with a questionnaire such as the "Motivated Strategies for Learning Questionnaire" (MSLQ) developed by Pintrich and DeGroot (1990). In this questionnaire, Pintrich and DeGroot identify three components of motivation toward academic performance. The first is a "self-efficacy" component, which reflects student's beliefs about their ability to perform the task. Second is an "intrinsic value" component, which reflects students' beliefs about the importance and interest of the task, and third is an "affective" component, which includes students' emotional reactions to the task.

Self efficacy was addressed by questions like "I expect to do very well in this class." These questions are thought to measure the students' expectations of success on the task. Intrinsic value is addressed by questions such as "I think I will be able to use what I learn in other classes." These questions are thought to measure one aspect of the student's expected reward for expending effort on the task. The affective component measures how students feel about the task. It was measured by Pintrich and DeGroot using "Test anxiety" questions (i.e. "I worry a great deal about tests"). Test anxiety was thought to be one of the most important affective reactions in a school context.

In addition to the three categories of motivation question mentioned above, Pintrich and

DeGroot included several questions about self regulated learning behavior. In particular, they asked about students' self-regulation strategies using questions like "I find that when the teacher is talking I think of other things and don't really listen to what is being said."

Using this instrument, Pintrich and DeGroot found that self-regulation, self-efficacy and test-anxiety were significant predictors of classroom performance. They also found that the motivational and self-regulated learning questions were significantly correlated with one another.

Other work has shown that motivation and interest can also affect text processing. For example, McDaniel et al. (2000) found that, compared to low-interest stories, high-interest stories required fewer attentional resources for processing, as measured by a secondary reaction time task. This was presumed to be because students were expending fewer attentional resources attempting to maintain concentration on the primary reading task. Importantly, McDaniel et al. found evidence that story interest was affecting the depth of text processing being done, with more interesting stories being processed more deeply.

A similar result was found by Schiefele (1996). In that study high-interest readers developed a deeper representation of the text's meaning than did low interest readers. Subjects were tested using a "recognition and verification" task, in which they had to decide whether a series of sentences had been present in the original text. The sentences represented exact matches, paraphrases, correct inferences or incorrect inferences from the original in such a way that verbatim, propositional and situation model representations could be assessed separately. Schiefele found that low interest readers had better verbatim representations, but that high interest readers built better propositional representations. Surprisingly, no difference was found for the situation model representation.

In a similar study, Schiefele and Krapp (1996) found that topic interest was significantly correlated with indicators of deep processing in a free recall task. Subjects with high interest recalled more idea units, elaborations and main ideas than subjects with low interest. Interestingly, the effect of topic interest was not related to measures of intelligence or prior knowledge.

A study by Boscolo and Mason (2003) also has particular relevance to our current work. Boscolo and Mason investigated the interaction between topic knowledge, textual coherence

and reader's interest. They were motivated by many of the same results discussed in Section 2.4, which show that textual cohesion can affect depth of processing, and by some of the work reviewed above showing that level of interest can also affect depth of processing. Students were randomized by level of interest and prior knowledge, and assigned to one of three texts that varied by level of coherence. After reading, students were given tasks which assessed their text base and situation model understanding. Results indicated that performance on all tasks improved with increasing interest, and also with increasing prior knowledge. They also found that the effects seemed to be additive, so that students with high knowledge *and* high interest did significantly better than the other groups. Surprisingly, they did not find the "cohesion reversal effect" interaction between knowledge level and textual cohesion.

These studies and others suggest that topic interest can have a similar effect to prior knowledge, leading to deeper processing and greater recall of a text.

Because these studies have suggested that student motivation is important in tutoring and also in text comprehension, I include a motivational survey in the current work. This survey is modeled on Pintrich and DeGroot's MSLQ. It is described more fully in Section 4.4.3 and evaluated in Section 5.1.1. The expected effect of motivation is described in Hypothesis 5 of the next section.

2.7 HYPOTHESES

The literature discussed in Section 2 above suggests three primary and two secondary hypotheses. The three primary hypotheses concern learning gains, and their relation to reflection, textual cohesion and prior knowledge. The two secondary hypotheses are more concerned with the mechanisms underlying the first three sets of results, and use evidence from cognitive load measures to address the question of whether textual cohesion and motivation affect inference during reading.

Primary Hypotheses

1. **Abstractive Reflection improves learning.** Based on the research reviewed in Sections 2.2 and 2.3, I expect students in the reflective reading conditions to learn more than

those in the no-reflection control, particularly in "far-transfer" measures of learning. I will also look for interactions between reflection and motivation.

2. **Reflective Cohesion Affects Learning.** Based on the research reviewed in Section 2.4, I expect that the cohesiveness of the reflective reading will affect its superiority over a no-reflection control. For most students, I expect high cohesion text to beat control by a larger margin than low cohesion text. I will also look for interactions between textual cohesion and motivation.
3. **The Impact of Reflective Cohesion on Learning Interacts with Knowledge.** Also based on the work in Section 2.4, I expect that the cohesiveness of the reflective reading will interact with student knowledge. I expect subjects with high pre-test scores to learn more from *low* cohesion reflective text, and subjects with low pre-test scores to learn more from *high* cohesion text. I will also look for interactions between knowledge, cohesion and motivation.

Secondary Hypotheses

4. **Textual cohesion affects learning through inference.** If the aptitude treatment interaction in Hypothesis 3 is confirmed, I expect that subjects who learn more from low cohesion text will also have higher cognitive load when reading that text than when reading the high cohesion reflective text. This hypothesis follows McNamara's explanation which was described at the end of Section 2.4. Higher cognitive load would be evidence that the gaps in the low cohesion text are triggering inference, which is then causing learning.
5. **Motivation affects inference.** Engagement with the reflective text may vary by motivation level. Based on the research described in Section 2.6, I expect higher levels of motivation to be accompanied by evidence of greater inference while reading the reflective text

3.0 EXPERIMENTAL TESTBED

In this chapter I describe the preexisting software used for this experiment. In Section 3.1 I describe the Itspoke tutor. I will briefly explain the origins and features of the version of Itspoke used in this study, and give a detailed description of the physics problems it tutors.

Then in Section 3.2 I describe Linger, the software used to collect the reading time and self-reported reading difficulty measures.

3.1 THE ITSPOKE TUTOR

Itspoke (**I**ntelligent **T**utoring **SPOKE**n dialog system) is a spoken dialog tutoring system which teaches five problems in qualitative physics. At the start of each tutoring session, a problem statement is presented at the top of the screen (see the screen shot in Figure 1 on Page 24). The tutor then asks a question, to which the student responds. The tutor walks through the correct solution to the problem, entering sub-dialogs as necessary to remediate student incorrect answers. The student wears a head-set, and so can hear the tutor's spoken utterance as well as read it on screen, but the student makes only spoken responses.

Itspoke was built by adding a spoken dialog interface to the Why2-Atlas tutoring system (VanLehn et al., 2002). Why2-Atlas is a *text based* intelligent tutoring system which interacts with the student using typed natural language dialog. Tutor questions and responses are determined using a finite state dialog manager (Rosé et al., 2001), which selects the next tutor utterance based on a semantic analysis (Rosé, 2000) of the previous student answer. Why2-Atlas also includes a system for essay analysis using the LCFlex parser (Rosé, 2000), but the version of Itspoke used in this study does not ask the student to submit essay

answers, so Why2-Atlas's essay functionality was not used. Itspoke uses speech recognition and text-to-speech software to convert Why2-Atlas's text interactions to and from speech. The version of Itspoke used in this study is identical to the control condition of Forbes-Riley and Litman (In press.). Besides removing the essay, this version differs from the original Why2-Atlas/Itspoke systems in that it has been reimplemented using the TuTalk (Jordan et al., 2007) dialog system. In addition, in this version only the tutor utterances are displayed.

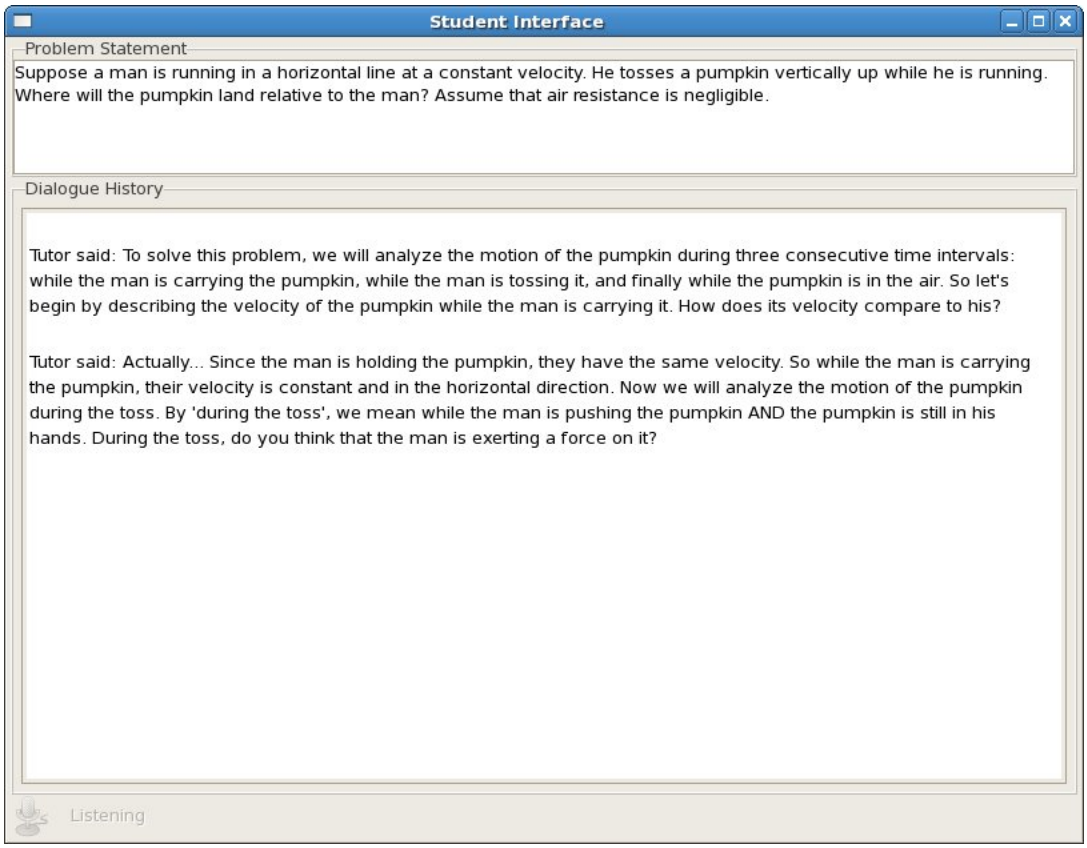

Figure 1: The Itspoke user interface

Why2-Atlas has an extensive evaluation history, and has been tested in several configurations. In experiment 1 of VanLehn et al. (2007) a version of Why2-Atlas was fielded which tutored ten problems in qualitative physics. In other experiments (e.g., exp. 5 of VanLehn et al. (2007)) a reduced set of only five of the problems was used, so the pre- and post-test questions were also reduced to include only the questions more closely related to the tutored material. Instead of 40 questions, that study used only 26 questions.

These five problems and their solutions (as embodied in Why2-Atlas's original "ideal essay") are briefly summarized below.

- **Elevator-Keys**

  Problem statement: "Suppose a man is in a freefalling elevator that has nothing touching it (you should ignore air resistance). The man is holding his keys motionless right in front of his face. He then lets go. He doesn't toss them up or throw them down; he just releases his grip on them. What will be the position of the keys relative to the man's face as time passes?"

  Solution Steps: "Since the man is holding the keys motionless while falling, the man and the keys must have the same velocity when he drops them. The elevator is falling freely. Consequently, everything in it, including the man and the keys, are falling freely vertically down. The man and the keys will, therefore, have the same free-fall acceleration g. Since the man and the keys have the same initial velocity and the same acceleration, both will have the same velocity at all times. Consequently, both have the same displacement from the point of release at any time during the fall. Thus, the keys will remain in front of the man's face at all times throughout the fall."

- **Plane-packet**

  Problem statement: "An airplane flying horizontally drops a packet when it's directly above the center of an empty swimming pool, where a bright red target is painted. Does the packet hit the target? Assume air resistance is negligible."

  Solution Steps: "The plane has a horizontal velocity when it drops the packet. The packet will have the same velocity as the plane at the instant of the drop. Thus at the instant of the drop, the packet has a horizontal velocity component (equal to the plane's horizontal velocity) and a vertical velocity component of zero. Neglecting air resistance, earth's gravity is the only force on the packet after it is released. This force is in the vertical direction and causes a downward vertical acceleration of the packet. Thus, the vertical component of its velocity keeps increasing, and the packet hits the ground after some time. During this period of time, the packet keeps traveling along the horizontal due to its horizontal velocity component. Therefore, it has a horizontal displacement

from the point of release when it hits the ground. Since the packet was dropped right above the target, it will miss the target by the amount of its horizontal displacement."

- **Earth-Sun**

  Problem statement: "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?"

  Solution Steps: "The force of gravity between the earth and the sun results from their interaction resulting from the gravitational pull between them. According to Newton's third law of motion, if a force is exerted by one body on a second body, the second body must exert a force of equal magnitude and opposite direction on the first body. This law is universal and is obeyed whenever a force is exerted. The sun, therefore, must experience a force of gravity due to the earth that is equal in magnitude and opposite in direction to the force of gravity on the earth due to the sun. "

- **Pumpkin**

  Problem statement: "Suppose a man is running in a horizontal line at a constant velocity. He tosses a pumpkin vertically up while he is running. Where will the pumpkin land relative to the man? Assume that air resistance is negligible."

  Solution Steps: "The man is holding the pumpkin and running with a constant horizontal velocity. In order to throw the pumpkin vertically up, he exerts an upward vertical force on the pumpkin. The downward force of gravity is always present but the force applied by the man is greater such that there is a net upward vertical force on the pumpkin while being thrown. This net force causes the pumpkin to accelerate upward and acquire an upward vertical velocity at the end of the throw. After the release, the only force acting on the pumpkin is the downward vertical force of the earth's gravity. This force decelerates the upward velocity of the pumpkin, eventually bringing it to zero, at which time the pumpkin begins to fall. The acceleration due to gravity is ever present, hence the downward velocity of the falling pumpkin keeps increasing until it lands. Since the vertical force of gravity is the only force acting on the pumpkin after the throw, there is no force acting on it in the horizontal direction. Thus, the pumpkin's horizontal velocity component, which is equal to the man's velocity at the time of throwing it, will remain constant. He continues to run with the same constant horizontal velocity, so the

pumpkin and the man have the same velocity along the horizontal at all times during the flight of the pumpkin. Consequently, the man and the pumpkin have the same displacement along the horizontal at all times during the flight of the pumpkin. Therefore, the pumpkin will land in the man's hands when it reaches that level."

- **Car-Truck**

  Problem statement: "Suppose a lightweight car and a massive truck are driving towards each other in a straight line. They hit a patch of frictionless ice and have a head-on collision. Upon which vehicle is the impact force greater? Which vehicle undergoes the greater change in its motion? (As usual, assume air resistance is negligible)".

  Solution Steps: "The car and the truck will both experience a force due to the impact. The force experienced by each vehicle results from the interaction of touch between them. These two forces resulting from the interaction are the action/reaction pair of Newton's third law of motion. In terms of the third law, the force on the truck exerted by the car and the force on the car exerted by the truck will be equal in magnitude and opposite in direction. According to Newton's second law of motion, the acceleration of a body is equal to the force acting on it divided by its mass. Since the magnitude of the forces of impact on both the vehicles are the same, the acceleration of the car will be much greater than that of the much more massive truck. A larger magnitude of acceleration implies a larger rate of change of velocity, which means greater change in motion. Therefore, the car undergoes a greater change in its motion."

In early studies, Why2-Atlas was used to evaluate the effectiveness of a typed natural language interface in tutoring, and to compare learning gains from tutoring to those from reading an expository text. In later studies, Itspoke was used to evaluate the effect of porting Why2-Atlas' typed interface to a spoken one. In the current study, Itspoke is used to provide tutoring over five problems in qualitative physics, which are then (in the experimental conditions) compared with each other in the post-tutoring reading.

3.2 THE LINGER INTERFACE

All readings were presented using the Linger (Rohde, 2003)"tap-to-read" interface. Linger is a platform for language processing experiments which presents text in incremental units (an increment can be at the word, sentence or text-block level), and records the reading time for each unit.

The Linger interface is shown Figure 2 presenting a block of text from one of the reflective readings. After reading this block, the student will tap a key and Linger will present the next block of text. Linger records the time between presenting the text and the next tap-for-more in milliseconds.

In addition, Linger allows a variety of questions to be asked following each text presentation. In the current study, I used Lingers likert scale implementation to ask students how hard or easy the previous text had been to read. The Linger interface is shown presenting this question in Figure 3.

Both of these measures were used to estimate cognitive load, as will be described in Section 4.4.2.

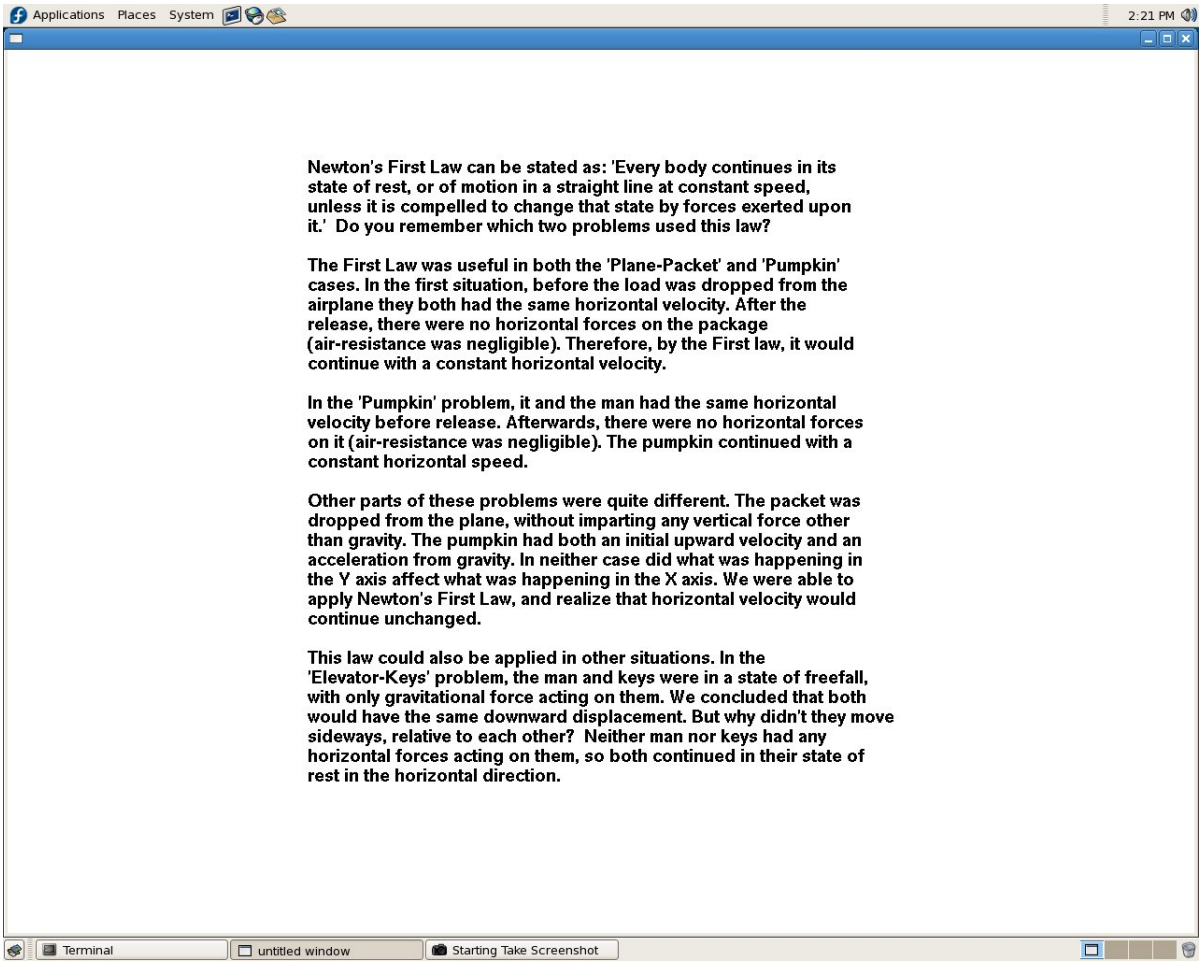

Figure 2: Linger Tap-to-read Interface

Figure 3: Linger difficulty rating screen

4.0 STUDY DESIGN

In this section I describe the experimental procedure, materials and measures used in this study. Section 4.1 provides an overview of the experimental conditions and procedure. Section 4.2 describes how participants were recruited and selected. Section 4.3 describes the reflective texts developed for the study. Section 4.4 describes the measures of learning and of cognitive load, as well as the motivational survey used for the study. Finally, Section 4.5 describes the statistical methods used for analysis.

4.1 OVERVIEW

Table 1 on Page 32 outlines the experimental procedure designed to test the hypotheses just described in Section 2.7. All subjects read and signed a consent form, provided background information such as high school GPA and SAT scores, read a non-physics "warm-up" reading, then read introductory material about physics. After reading the introductory material, they took a pre-test to measure their domain knowledge before tutoring. The pre- and post-tests are described more fully in Section 4.4.1. Each subject's pre-test score was compared to the distribution of scores on similar tests used in previous studies. Subjects whose scores fell in the middle third of the expected range were dismissed at this point. Subjects who fell in the upper or lower third continued to the next phase of the study, and were recorded as "high" and "low" pre-testers, respectively. Extreme groups design is described more fully in Section 4.2.

Next, the subjects took a motivational survey (described in Section 4.4.3), then engaged the Itspoke system (described in Section 3.1) in interactive tutorial dialogs which covered

five physics problems. The Itspoke system was identical for all subjects. After tutoring there was a second motivational survey, which was identical to the first except for minor changes in tense. After this the subjects read a post-tutoring text which varied by condition. After the post-tutoring text was a post-test, followed one week later by a delayed post-test. Retention testing with a 1 week delay has also been used in studies with the Why2-Atlas tutor (VanLehn et al., 2007).

The study contained a control condition and two experimental conditions. In the control condition, subjects re-read a shortened version of the introductory physics text (the "again" condition). In the *high* cohesion reflection condition ("refHigh"), subjects read a reflective text with high cohesion, and in the *low* cohesion reflective condition ("refLow"), subjects read a reflective text with low cohesion. These texts will be described more fully in Section 4.3.1.

Control ("again")	**High Cohesion ("refHigh")**	**Low Cohesion ("refLow")**
non-physics warmup reading	non-physics warmup reading	non-physics warmup reading
pre-reading	pre-reading	pre-reading
pre-test	pre-test	pre-test
Motivation Survey	Motivation Survey	Motivation Survey
tutoring dialogs	tutoring dialogs	tutoring dialogs
Motivation Survey	Motivation Survey	Motivation Survey
Shortened Pre-reading	**Reflective Reading HIGH Cohesion**	**Reflective Reading LOW Cohesion**
Post-test	Post-test	Post-test
Delayed Post-test	Delayed Post-test	Delayed Post-test

Table 1: Study Design. The portion that differs between conditions is shown in bold.

All readings were presented on screen using Linger's "tap-to-read" interface, as shown in Section 3.2. Linger also collected the motivation survey. The pre- and post-tests were administered using a web interface, and their results recorded in a relational database.

Hypothesis one, that adding a reflective reading improves learning, was tested by combining the two reflective reading conditions into one "ref" condition, and comparing its learning gains to those of the control "again" condition. I expected that this combined reflective

reading condition would show larger learning gains than the control.

Hypothesis two, that reflective cohesion affects learning, was tested by comparing learning for all students between all three experimental conditions: "again, refHigh or refLow". I expected students in the high cohesion reflective condition to beat the control by a higher margin than those in the low cohesion reflective condition.

Hypothesis three, that reflective cohesion interacts with knowledge, was tested by comparing learning gains of high vs low knowledge subjects in the two reflective reading conditions. I expected that high knowledge subjects would learn more from the low cohesion reflective text, and that low knowledge subjects would learn more from the high cohesion reflective text. The High and Low knowledge groups ("hiPre" and "loPre" respectively) were determined based on pre-test scores as described in Section 4.2.

Hypothesis four was that the aptitude treatment interaction predicted in Hypothesis three would be caused by differing amounts of inference when reading the high vs low cohesion text. This hypothesis was tested by comparing the cognitive load measures described in Section 4.4.2. I expected students who learned more from the low cohesion text to also have higher cognitive load when reading that text than when reading the high cohesion text. Similarly, I expected the students who learned more from the high cohesion text to have higher cognitive load when reading that text than when reading the low cohesion text.

Hypothesis five, that student motivation affects inference from text, was tested by comparing measures of cognitive load between different levels of student motivation. I expected more highly motivated students to engage the text more actively, to make more inferences, and so have a higher cognitive load. In addition, I investigated using motivation level as a covariate when testing the primary hypotheses, above.

4.2 PARTICIPANTS

Because hypotheses two and three involved comparing learning gains between high and low pre-testers, our study used an "extreme groups" design (Feldt, 1961), to increase the power of this comparison. Extreme groups design is very common in the psychological literature,

and was also used in the Boscolo and Mason study (Boscolo and Mason, 2003) described in Section 2.

I established high and low thresholds by examining data collected from previous studies with the Itspoke tutor in 2005 and 2008. Thresholds were designed to divide the subject pool into roughly even thirds, assuming the distribution of pre-test scores would be the same as in previous years.

After each student's pre-test, a selection score was calculated based on the subset of 26 questions which had also been used in 2005 and 2008 (the 26Q "near transfer" questions). Subjects whose selection scores fell between the high and low thresholds were paid (or given course credit) and dismissed without moving on to tutoring. These thresholds produced fairly even splits on the new data.

In total 166 students were recruited by flyer, by advertisement during an undergraduate psychology course, and by inclusion in the psychology subject pool. Of those 166, 40 were dismissed as middle-third pre-testers and not used in the study. An additional 12 were removed because they did not return for the delayed post test, 9 were removed because they did not provide complete background information (missing SAT scores or high school gpa), 6 were removed because of incomplete or missing pre-tutoring motivation surveys. This left a total corpus of 99 subjects of which 45 were low pre-testers and 54 were high pre-testers. Subjects were assigned randomly to experimental conditions, which resulted in an allocation of 33 subjects to the control condition, 32 to the high cohesion condition, and 34 to the low cohesion condition. 47 subjects were paid for their participation and the remainder were taken from the psychology subject pool. The distribution of subjects among categories and conditions is shown in Tables 14, 15 and 16 on page 57.

	Median	**Mean**	**N**
Low pre-test	.385	.374	45
High pre-test	.692	.682	54

Table 2: Pre-test scores for high and low pre-testers

Table 2 shows the median and mean pre-test scores on the 26Q subset for the accepted high and low pre-testers. A two sided t-test shows that pre-test scores were significantly

different between pre-test groups both for the 26Q subset used for selection ($p < 2.2e\text{-}16$) and also for the full set of 44 questions ($p < 2.2e\text{-}16$).

4.3 MATERIALS

4.3.1 Reflective Texts

Section 2.2 reviewed literature which suggested that one key to transfer was creating a more abstract schematic representation of the learned material. The literature also suggested that an effective way to create such an abstract representation would be to explicitly compare two or more analogs, noticing which features were common between them and which were incidental surface details. Section 2.3 reviewed literature from reflection research, which suggested that reflection after problem solving was a good way to generate these abstracted representations. It suggested in particular helping students derive abstractions about the problem solving process by comparing multiple problem solving performances simultaneously.

Based on these considerations, I added reflective readings to the Itspoke tutor. In general, the readings concentrated on places in which the same physics principle had been applied in different situations, and on similarities in the overall problem solving approach between problems. The readings were structured to roughly follow the four steps described by Tchetagni et al. (2007), which were mentioned in Section 2.3.2. They first elicited curiosity by asking a question about the similarities of the tutored problems. They then reviewed relevant parts of several of the problems, pointing out which were common and essential, and which were unimportant. Finally, they (where possible) "evaluated" the correctness of the commonalities derived by showing how they would work in another of the tutored problems.

The rest of this section demonstrates how a portion of one reflective text was developed, based on how Newton's third law is discussed in two of the Itspoke tutored problems. In the Earth-Sun problem Newton's third law is used to motivate the idea that the earth pulls on the sun with exactly the same force as that with which the sun pulls on the earth (but in the opposite direction). In the car-truck problem, it is used to motivate the idea that the force

of the car hitting the truck is the same as that of the truck hitting the car. The following excerpts show how the tutor describes these points in two example dialogs:

> **From Earth-Sun:** Okay. Newton's third law says every force has an equal and opposite reaction force. That is, if there is a force acting on object A due to object B, then there is also a force acting on B due to A. These two forces have the same magnitudes but opposite directions. Moreover, they are the same type of force. If one is a gravitational force, then so is the other. If one is a frictional force, then so is the other. In this case, there is a gravitational force on the earth due to the sun. Is there a gravitational force on the sun due to the earth?

> **From Car-Truck:** Alright. Newton's third law says that every force has an equal and opposite reaction force. That is, if there is a force acting on object A due to object B, then there is also a force acting on B due to A. The two forces have the same magnitude and opposite directions. So in our problem, upon which vehicle is the impact force greater?

These points were compared in the reflective reading by first reminding the student that they had occurred in the dialogs, then rhetorically asking what they had in common. The reading then pointed out that they both required the use of Newton's third law to find action/reaction pairs. Then the cases were contrasted by pointing out that the forces involved operate along different axes (vertical in earth-sun and horizontal in car-truck). Another point of contrast was that that while they were different forces in each case (*gravity* in the case of earth-sun and *impact* in the case of car-truck) the important point is that both forces in the same action/reaction pair have to be the same type of force. The reading also reminded the student that the point of finding an action/reaction pair in both cases was to show that their magnitudes were equal. Finally, the reading "evaluated" the correctness of these points by applying them in a third situation, the Plane-Packet problem. To the extent possible, each point in the reading was constructed this way.

For the reasons discussed in Section 2.4 on Page 13, both "high" and "low" cohesion versions of the reflective text were written. Both versions were written to contain similar rhetorical questions, and to have roughly the same semantic content and the same general structure, however the high cohesion version was constructed to contain fewer "inferential gaps."

The high and low cohesion versions were constructed using similar methods to those used by McNamara in generating her texts. Examples of her "heart disease" text are shown in the

Appendices. Appendix D.1 shows her high cohesion version, and Appendix D.2 shows her low cohesion version. The high cohesion version was written to remove places in which inference was required to understand the low cohesion text. For example, referring expressions were made explicit, and causal and logical relations that were only implied in the low cohesion version were spelled out. Examples of this process can be seen by comparing the first two paragraphs of each text.

Similar changes were made to make "high" and "low" cohesion versions of the reflective texts written for our experiment. The next two listings show high and low cohesion excerpts from the part of the reflective texts that discusses how Newton's Third Law was used in the tutored problems.

High cohesion excerpt: Newton's Third Law

Newton's Third Law

In the Car-truck problem we wanted to compare the relative accelerations of the car and truck. Therefore, we first had to compare the impact force of the car on the truck with the impact force of the truck on the car. Similarly, in the Earth-Sun problem we were asked to compare the force of the Sun's pull on the Earth with the force of the Earth's pull on the Sun. Do you remember which of Newton's Laws was useful in these two problems?

In these two problems we used Newton's Third Law to show that the forces involved in an action/reaction pair had the same magnitude but acted in opposite directions to each other. An action-reaction pair is formed whenever one object exerts a force on a second object. Newton's Third Law says that when one object exerts a force on a second object, there is an equal and opposite reaction force from the second object back onto the first object. In addition, the type of force is always the same for both objects in the action/reaction pair. For example it was gravitational force on both Earth and Sun, and impact force on both car and truck. The two forces in an action-reaction pair can operate along any axis, but always have opposite directions to each other. For example, the earth pulled in the opposite direction than did the sun (vertically down vs vertically up), and the car's impact force was opposite to the truck's (horizontally right vs horizontally left).

In both the Car-truck and Earth-Sun problems, using Newton's Third Law allowed us to see that the forces acting on each object in the action/reaction pair had the same magnitude, even though the objects in the pair had different masses. The Earth pulls as hard on the Sun as the Sun pulls on it, even though the Sun is more massive. Similarly, the car hit the truck with as much force as the truck hit it, even though the truck had more mass.

You can use the idea of an action/reaction pair to analyze any problem in which one object exerts a force on another object. For example in the Plane-Packet problem, the Earth exerts a gravitational force on the packet, and the packet accelerates downward toward the Earth. Does the Earth also accelerate toward the packet? Yes. Earth and packet form an action/reaction pair, linked by gravitational attraction. The packet pulls on the Earth with gravity as hard as the earth pulls on it. The Earth therefore accelerates toward the packet, although less noticeably because of its greater mass.

Low cohesion excerpt: Newton's Third Law

In the Earth-Sun problem we had to compare the strength of the Sun's pull on the Earth with that of the Earth's on the Sun. In the Car-Truck problem we had to compare the force of the car's impact on the truck with that of the truck on the car. Do you remember which of Newton's Laws was useful in these cases?

In both situations we used the Third Law to show that the forces involved in an action/reaction pair had the same magnitude but acted in opposite directions. An action-reaction pair is formed whenever one object exerts a force on another object. Newton's Third Law says this force will have an equal and opposite reaction force. The type of force is always the same for both objects in the pair. It was gravitational on both Earth and Sun, and impact on both car and truck. They can operate along any axis, but always have opposite directions to each other. For example, the earth pulled in the opposite direction of the sun (vertically up vs vertically down), and the car's impact force was opposite to the truck's (horizontally right vs horizontally left).

In both of these situations, using Newton's Third Law allowed us to see that the forces acting on each object in the action/reaction pair had the same magnitude, even though the objects in the pair had different masses. The Earth pulls as hard on the Sun as the Sun does on it. The car hit the truck with as much force as the truck hit it, even though their masses were very different.

You can use the idea of an action/reaction pair to analyze any situation in which a force is exerted. After the plane releases it, the force of gravity becomes the net force, and the packet accelerates downward toward the Earth. Does the Earth accelerate toward the packet? Yes. Earth and packet form an action/reaction pair. The packet pulls on the Earth as hard as the earth pulls on it. The Earth therefore accelerates toward the packet, although less noticeably because of its greater mass.

Notice first that the high and low cohesion versions of the text are semantically very similar. Each point expressed in the high cohesion version is also expressed in the low cohesion version, but there are many differences in how the ideas are expressed. In general, the high cohesion version uses more consistent referring expressions and makes the relationships between ideas more explicit. For example, the high cohesion version uses "problem" throughout, while the low cohesion version also uses other referring expressions such as "case" or "situation." Also, the low cohesion excerpt opens by saying that we wanted to compare the strengths of the Sun and Earth's gravitational pull, but doesn't say why. The high cohesion version says why. The high cohesion version uses bridging expressions such as "similarly" or "for example" that make the relationship between adjacent ideas more explicit, while the low cohesion example forces the reader to infer these relationships. The high cohesion version also includes topic headings, to make the overall structure of the text more clear. In the low cohesion version, this structure had to be inferred.

These changes make the high cohesion versions longer than the low cohesion versions. The top four lines of Table 3 show word counts for four texts used by McNamara in her cohesion work. As can be seen, in each case the "low" cohesion version has a lower word count than the corresponding "high" cohesion version. On average, her low cohesion texts are only 73% as long as her high cohesion texts. The bottom rows of Table 3 show word counts for the high and low cohesion texts developed for this study. For these texts the low cohesion version is 71% as long as the high cohesion version, right in the middle of McNamara's range.

The read-again control, which is reproduced in the second part of Appendix A.1, had 2013 words. It was a shortened version of the introductory reading, from which some content had been removed to control for length relative to the reflective readings. Sections were selected for removal if they covered concepts not in the reflective readings, with the result that all the post-test readings covered a similar set of physics topics. Note that this process of reduction left the length of the read-again text right in the range between the high cohesion and low cohesion reflective text's word counts.

	McNamara Text Word Counts			
	Low	**High**	**diff**	**Low/High**
Cell Division	650	902	252	0.72
Air War	1036	1300	264	0.80
Heart Disease	682	1052	370	0.65
Traits of Mammals	590	817	227	0.72
	Reflective Reading Word Counts			
	Low	**High**	**diff**	**Low/High**
Qualitative Physics	1541	2161	620	0.71

Table 3: Comparing Word Counts: High vs Low Cohesion Texts

Having written "high" and "low" cohesion versions of the reflective texts I verified that their cohesiveness really varied in the intended way by using the CohMetrix tool (Graesser et al., 2004). CohMetrix produces a wealth of interesting output, but a great deal of it is not clearly relevant to my design goals. The full CohMetrix output is reproduced in Table 41 in Appendix B. Selected measures that seem most relevant are shown in Table 4 below.

My design goals in designing these texts were to vary the degree to which they used con-

Measure Description	Low Cohesion	High Cohesion
CREFC1u 'Prop. of content words that overlap between adj. sent.'	0.11	0.16
LSAassa 'LSA, Sentence to Sentence, adjacent, mean'	0.29	0.39
LSApssa 'LSA, sentences, all combinations, mean'	0.3	0.39
LSAppa 'LSA, Paragraph to Paragraph, mean'	0.41	0.51
CAUSVP 'Incidence of causal verbs, links, and particles'	54.53	56.38
CONADpi 'Incidence of positive additive connectives'	31.71	35.63
CONTPpi 'Incidence of positive temporal connectives'	8.24	9.92
CONCSpi 'Incidence of positive causal connectives'	15.22	20.75
CONLGpi 'Incidence of positive logical connectives'	18.39	26.61
CONLGni 'Incidence of negative logical connectives'	8.24	9.47

Table 4: Selected CohMetrix output for high and low cohesion physics texts

sistent expressions and the extent to which they made various logical and causal relationships explicit. The first row in Table 4 measures the extent to which content words are re-used in adjacent sentences. A higher number confirms that our high cohesion text is actually more cohesive along this dimension. Similarly, the LSA measures in rows 2, 3, and 4 measure how semantically similar adjacent portions of the text are. For each of these measures my high cohesion text is more cohesive than the low cohesion version. The six measures at the bottom of Table 4 measure the extent to which various types of relationship are spelled out in the text. For each of these measures, my high cohesion text again scores higher than the low cohesion text. In general, the CohMetrix tool seems to confirm that the cohesiveness of my high and low cohesion texts varied in the way we intended.

The full post-reading texts for the "again," "refHigh" and "refLow" conditions are reproduced in Appendices A.2, A.3 and A.4. The baseline and introductory pre-reading texts are reproduced in Appendix A.1.

4.4 MEASURES

4.4.1 Learning Measures

As described in Section 2.3, a major benefit from metacognitive activities such as abstraction and reflection may be in "far transfer" learning. To investigate the effect of reflection on far transfer we exploited the fact that different subsets of pre- and post-test questions had previously been created for use with different versions of the Why2-Atlas system.

Section 3.1 described how Itspoke implemented the five problems that had been used in a reduced version of Why2-Atlas. That version of Why2-Atlas had also removed 14 questions from the pre- and post-tests, only using 26 questions that were judged to be more related to the remaining problems. We use this 26 question (26Q) set as a measure of near transfer. To the 14 questions that had been removed, we add 4 new questions designed specifically to be dissimilar to the tutored problems, and use the resulting 18 question set (18Q) to measure far transfer learning. All of the questions involve using Newton's Laws or problem-solving principles to reason about force and motion problems, but they differ in how similar they are to the tutored problems.

In the 26Q "near transfer" set, surface features are often changed, but underlying problem situations are the same as in the tutored problems. For example as described on page 26 of Section 3.1, one of the tutored problems involves a man running at a constant horizontal velocity and throwing a pumpkin straight up. The student is expected to reason that the pumpkin has the same horizontal velocity as the man, and therefore the same horizontal displacement after the toss, and will land back in the man's hands. Compare that tutored problem with the situation described in the following question:

> A man is standing on a train that is moving horizontally in a straight line at a constant speed. He throws a steel ball straight up. Immediately after he releases the ball, what is the relationship between the horizontal speed of the man and the horizontal speed of the ball? (Assume air resistance is negligible)

This question is isomorphic to part of the pumpkin problem. A majority of the questions in the 26Q question set are isomorphic to a tutored problem in this way.

In contrast, the 18Q set tests the same physics principles as the "near transfer" questions,

but they are almost entirely non-isomorphic to the tutored problems. For example, one of the far transfer questions reads like this:

> The American bobsledders have a bobsled that weighs 300 N. The Jamaican's bobsled weighs 280 N. If the bobsledding teams are equally strong (i.e., they can push on the bobsled with the same force) and push the sled for the same period of time until they jump into it, which team's bobsled will be moving faster at the moment they jump into it?

This question tests physics concepts similar to those used in the tutored force-and-motion problems, however the structure of the question is different from any of the tutored problems. This question asks the student to use Newton's Second Law to reason that since the forces on the bobsleds are the same, the less massive bobsled will accelerate faster. Newton's Second Law is also used in several tutored problems, the most similar probably being the "car-truck" problem. In that problem it was used to deduce that the car would have a greater acceleration than the truck. Although the Second Law is used the same way, the many differences in the problem situation (push vs impact force, for example) are sufficient to classify this as a "far transfer" problem.

I use these question sets to measure "overall" learning (using the whole 44Q set), "near transfer" learning (using the 26Q set) and "far transfer" learning (using the 18Q set) for both immediate and delayed post-tests. A delayed post-test was also given to measure the long term retention aspect of robust learning, as described in Section 2. The immediate post-test contained questions that were isomorphic to those on the pre-test, while the delayed post-test contained the same questions as the pre-test. This gives us six measures of learning: "overall," "near," "far," "delayed overall," "delayed near," and "delayed far."

Each of these six measures is reported as Normalized Learning Gain (NLG). NLG is computed as (post-pre)/(1-pre) where "post" is percentage correct on the post-test and "pre" is percentage correct on the pre-test. NLG is an instance of a POMP (**P**ercentage **O**f **M**aximum **P**ossible achievement) score as described by Cohen et al. (1999) which expresses learning gains as a percentage of the amount remaining to be learned after the pre-test. Hake, when discussing the use of pre and post tests in physics education (Hake, 2007), argued that this measure of average normalized gain was "a much better indicator of the extent to which a treatment is effective than is either gain or posttest" (pg 9). Perhaps for this reason, NLG

is a commonly reported measure in tutoring research (e.g.: (Graesser et al., 2003) [1].

The full set of post-test questions is reproduced in Appendix C. For reference, the questions in the 18Q set are marked "18Q(far)."

4.4.2 Cognitive Load Measures

As mentioned in Section 3.2, the readings were presented in a series of text blocks. Reading time was recorded for each block using Linger, and after each block the student was presented with a 1 to 7 likert scale asking how hard the block had been to read.

Linger presents a wide range of options in how to present textual material. It can be presented word by word, line by line or block by block. In addition the previous words or lines presented can disappear or remain on screen after the next ones are presented. We chose among these options using an informal usability study.

A major design goal was to make data collection as unobtrusive as possible, and to erect minimum barriers to text comprehension. Presenting text word-by-word or line-by-line made it extremely hard to comprehend. Similarly, using smaller block sizes made comprehension difficult, because when reading certain topics the original referent would often disappear before the reader had finished its explanation.

After experimentation, it was decided to size blocks of text to match individual topics as much as possible. The number of blocks per text and mean number of words per block are presented in Table 5. Note that the median block size for the high cohesion reflective text is larger than for the low cohesion text. The high and low cohesion texts had the same topic structure and therefore the same number of blocks, with the exception that one of the longer blocks had to be split between two screens in the high coherence version, resulting in one extra block of text.

I first describe how the self-report "hardness" ratings collected after each block were turned into cognitive load measures. The per-block self-reports were converted by first taking the mean of the responses over all the text blocks in a reading. Separate means were calculated for the non-physics "warm up" reading, for the introductory physics reading,

[1] See also (Jackson et al., 2004) where this measure is called an "estes score," and the term "normalized learning gain" is used for a different measure ($post - pre/SD_{pre}$).

Text	Number of Blocks	Mean Words per Block
non-Physics warmup	3	190.7
pre-reading	15	135.1
again	11	184.7
refLow	6	260.3
refHigh	7	273.1

Table 5: Mean block sizes for each text.

and for the post-tutoring reading. This resulted in three hardness ratings for each student. Measures were also calculated using the median rather than the mean.

Using these measures in their raw form carried the risk that there might be some bias in self-reporting between subjects, beyond differences caused by text difficulty. In other words, some subjects might just tend to rate all texts harder than other subjects. I corrected for this by subtracting the rating on the target text from the rating on a baseline text, with the baseline selected according to the hypothesis being tested. The underlying difficulty rating is on a 1-to-7 scale, with 1 being "extremely easy" and 7 being "extremely hard." So, if the post-tutoring reading is generally judged to be harder than the baseline this number will be positive. A higher number on the self-report measure indicates a higher cognitive load.

In Section 5.6.1 I use this self-report measure to test the hypothesis that differences in textual cohesion affect cognitive load. Because the important variable in this comparison was the text's cohesion, rather than its content, I used the physics pre-reading for the baseline measure.

The second measure of cognitive load was based on per-block reading speeds. Linger's tap-to-read interface presented blocks of text, and recorded the number of milliseconds between when the student was presented the text and tapped for the next block. The elapsed time was divided by the number of words in the block to arrive at a words-per-minute (WPM) reading speed. I again averaged the reading speeds over all the blocks in a text to arrive at a mean reading speed for each text. Separate means were calculated for the non-physics "warm up" reading, for the introductory physics reading, and for the post-tutoring reading.

This resulted in three reading speed ratings for each student. Measures were also calculated using the median rather than the mean.

As with the self-reports, using these measures in their raw form carried the risk that there might be some bias in reading speed between subjects, beyond that caused by text difficulty. I corrected for this by subtracting the speed on the target text from the speed on a baseline text, with the baseline selected according to the hypothesis being tested. Following the literature (i.e. (Schultheis and Jameson, 2004)), I assume that students read easy texts faster than hard texts. Therefore, a higher number on the baseline-corrected reading speed measure indicates a lower cognitive load. For example, assume a student read the non-physics baseline text at 200 words per minute (WPM). After tutoring, this student read the post-tutoring reading at 300 wpm (students often read the final reading more rapidly than the first one, possibly because it was the third time they had been exposed to the material: first in the physics pre-reading, second during tutoring, and third from the post-tutoring reading). The difference in reading speed for this student would be 300-200 = 100 wpm. Now consider a second student whose difference in reading speed was 300-250 = 50 wpm. This student showed a smaller increase in reading speed relative to the baseline, which suggests a relatively higher cognitive load than the first student.

Similarly to the self-report measure, the baseline measure was selected according to the hypothesis being tested. In section 5.6.2 I use the reading-speed measure to test the hypothesis that motivation affects cognitive load when reading. Because the important variable in this comparison is motivation toward physics, I use the non-physics warm-up reading as a baseline. Subtracting the baseline adjusts for individual differences in reading speed, so that the resulting number shows more clearly the effect of any difference in topic motivation.

4.4.3 Motivational Survey

The potential effects of motivation on text processing and learning were reviewed in Section 2. In this section I describe the motivation instrument used in this study. Section 5.1.1 validates the instrument on our collected data.

As mentioned in Section 2, Pintrich and DeGroot (1990) developed the "Motivated Strategies for Learning Questionnaire (MSLQ)" for measuring motivation. The MSLQ includes questions which measure, among other things, the students' self-regulation behavior, attitudes about self-efficacy, and beliefs about the intrinsic value of the task. In this work I use a reduced version of the MSLQ, which is also patterned on an instrument used in previous work by Ido Roll (Roll, 2009)[2]. The instrument used in this study is shown in Figure 4.

Please read the following statements and then click a number on the scale that best matches how true it is of you. 1 means "not at all true of me" whereas 7 means "very true of me".

1. I think that when the tutor is talking I will be thinking of other things and won't really listen to what is being said.
2. If I could take as much time as I want, I would spend a lot of time on physics tutoring sessions.
3. I think I am going to find the physics tutor activities difficult.
4. I think I will be able to use what I learn in the physics tutor sessions in my other classes.
5. I think that what I will learn in the physics tutor sessions is useful for me to know.

Figure 4: Pre-tutoring Motivational Survey

Questions one and two address self-regulation, particularly the students' tendency to manage and control their own effort. Question one is on a reversed scale relative to the other questions, so responses to it were inverted. Question three addresses self-efficacy, the students expectation of success on the task. Questions four and five address intrinsic value, the student's beliefs about the importance and interest of the task.

Although these different types of motivation question are theoretically distinct, we will find in Section 5.1.1 that their responses were highly correlated with each other in our corpus.

[2]I am very grateful to Maxine Eskenazi for providing the survey used in this study.

4.5 STATISTICAL METHODS

The major statistical method used in this study is the ANOVA (**AN**alysis **O**f **VA**riance). In its simplest form, an Anova tests the null hypothesis that the means of two or more groups are the same. It does this by comparing the variance *within* groups (the error variance) to the variance *between* groups (the effect variance). The ratio of these two variances is called the F-measure. The higher the between group variance is relative to the within group variance, ie: the higher the F-measure, the more likely it is that there is a "true" difference between the group means.

When we examine differences in the levels of only one explanatory variable, or factor, it is called a "one factor design." A one factor analysis is used in the current study to compare different values of the factor "experimental condition," which as described in Section 4.1 either "again, refHigh and refLow" or "noRef, ref." This will be used to test hypotheses one and two, as shown in the top two rows of Table 6 .

When more than two levels of an explanatory factor are being compared, the Anova will indicate the probability that there is a difference between group means, but will not tell us which means are different from one another. In these situations, we use the post-hoc TukeyHSD (Honest Significant Difference) test. This test performs multiple pairwise comparisons between the means and reports which differences are significant after correcting for multiple comparisons.

When comparing the effects of more than one explanatory factor, we have a "multi-factor design." This design raises the possibility of interactions between the explanatory variables. For example, to test hypothesis three in the current study we simultaneously compare the effects of both pre-test category and cohesion condition on learning. This is shown in the third row of Table 6. In general, a significant interaction between two factors indicates that the effect of one factor on the outcome variable (which in our case is normalized learning gain or cognitive load as described in Sections 4.4.1 and 4.4.2) is different for different values of the other factor. We interpret significant two-way interactions by examining the mean normalized learning gains for each combination of factors in the 2x2 design.

In this work we also encounter three-way interactions between the effects of three dif-

ferent independent variables: motivation, pre-test category and experimental condition. We interpret these using one of the methods described by Roberts and Russo (1999, p. 212). As Roberts and Russo note, a three way interaction means that there are "different two way interactions between two of the factors according to the value of the third factor." We analyze these by splitting the data according to the levels of the third factor, and performing two way Anovas on the remaining two variables. This method is easy to interpret, however splitting the data results in reduced degrees of freedom, and a correspondingly more stringent F-measure. Our results, described in Section 5, are strong enough to survive the slightly reduced sensitivity.

In Section 5.3.2 I will also use this method to interpret a two way interaction. I will split the data by the second predictor in order to show that the first predictor is only significant at certain values of the second.

In each Anova described in the top of Table 6, the dependant measure is Normalized Learning Gain on either "overall," "near," "far," "delayed overall," "delayed near," and "delayed far" questions (see Section 4.4.1). The independent measures are experimental condition ("again- refHigh- refLow" or "noRef -ref."), pre-test category ("highPre-loPre") and motivation level ("lowMot, midMot, hiMot").

The statistical methods used to test secondary hypotheses four and five are identical to those used to the first three hypotheses, with the exception that the dependant variable is cognitive load, rather than normalized learning gain. This is shown in the bottom two rows of Table 6.

All statistics were calculated using the R statistical language (R Development Core Team, 2005).

	Hypotheses	Outcome Variable	Independent Factors
Primary Hypotheses			
1	Abstractive Reflection improves learning	NLG	Reflective Category (ref, noRef)
2	Reflective Cohesion Affects Learning	NLG	Exp. Condition (again, refHigh, refLow)
3	The Impact of Reflective Cohesion on Learning Interacts with Knowledge	NLG	Reflective Cohesion (refHigh refLow) Pre-test category (highPre, LowPre)
Secondary Hypotheses			
4	Textual cohesion affects learning through inference	Cog. Load	Reflective Cohesion (refHigh refLow) Pre-test category (highPre, LowPre)
5	Motivation affects inference	Cog. Load	Motivation (lowMot, midMot, hiMot)

Table 6: Outcome Variables and Factors per Hypothesis

5.0 RESULTS

In this section I first examine whether the motivation instrument, the "near-far" division of questions and the cognitive load measures behave sensibly on the collected data. Following that, I test each of the hypotheses described in Section 2.7.

The distribution of scores is shown in Table 7. Note that no students answered all questions correctly, and that the mean scores were well below the maximum possible score (44). Similarly, no students answered all problems incorrectly. This suggests that there are no ceiling or floor effects in this data set.

	Min.	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**
pre-test	14.00	20.50	26.00	25.82	30.50	42.00
post-test	19.00	28.00	34.00	32.46	37.00	41.00
delayed post-test	16.00	26.00	31.00	31.12	36.00	43.00

Table 7: Correctness distribution on pre-, post- and delayed post-tests. N = 99

5.1 EVALUATING MEASURES

In this section I use the collected data to evaluate the motivation and cognitive load measures used, and also to evaluate the division of test questions into "near" and "far" subsets.

5.1.1 Evaluating the Motivation Instrument

After data collection, I used correlation analysis and Cronbach's Alpha (Cronbach, 1951) to explore whether the various questions in the motivational survey were measuring distinct

aspects of motivation, as formulated by Pintrich and DeGroot.

	Q1	**Q2**	**Q3**	**Q4**	**Q5**
Q1	1.000	0.223*	-0.295**	0.309**	0.459**
Q2	0.223*	1.000	-0.040	0.349**	0.459**
Q3	-0.295**	-0.040	1.000	-0.134	-0.088
Q4	0.309**	0.349**	-0.134	1.000	0.519**
Q5	0.459**	0.459**	-0.088	0.519**	1.000

Table 8: Correlation matrix for motivation responses. ** = $p < .01$, * = $p < .05$

Table 8 shows all the pairwise correlations between question responses in our corpus. Note that most of the questions are significantly correlated with all the other questions, with the exception of question 3, the self-efficacy question. This suggests that, with the exception of question 3, our questions are not distinguishing between different aspects of motivation.

Next we use Cronbach's Alpha (Cronbach, 1951) to assess the reliability of the measure. Cronbach's Alpha measures the internal consistency of responses to a multi-point questionnaire (i.e. 1 to N likert scales). Alpha is a function of the number of test items and the average correlation between them. Higher values are thought to indicate that the various test items are measuring the same underlying latent construct.

A common rule of thumb in psychological research is that alphas in the .6 to .7 (e.g. Gliem and Gliem (2003)) range indicate acceptable reliability for an instrument. Cortina (1993) acknowledges this practice, but cautions that the number of items on the instrument should be taken into account. Cortina shows that an alpha of .8 can be produced by a 3 item scale with average interitem correlations of .57. However, the same alpha can be produced with average interitem correlations of only .28 if the scale has 10 items.

Table 9 shows Alpha scores for various subsets of the motivation questions. The first row shows an Alpha of .53 for the full set of five questions. This is slightly lower than the generally accepted range discussed above. The second row shows Alpha after removing question 3 (the most poorly correlated question, as shown in Table 8). With question 3 removed Alpha rises to .71. For comparison, the bottom rows of Table 9 show Alpha for other subgroups of questions. Questions 4 and 5, the "intrinsic value" questions have an

alpha of .68. Adding question 2 improves this to .7, however the highest Alpha results from the set of questions 1,2,4 and 5, as shown in row 2. For this reason we remove responses to question 3 in the analysis below.

Questions	Alpha
1, 2, 3, 4, 5	0.531
1, 2, 4, 5	0.716
4, 5	0.683
2, 4, 5	0.703

Table 9: Cronbach's Alpha for subsets of motivation responses

As mentioned in Chapter 4, motivation was evaluated twice: once before tutoring and once after. Mean motivation was 4.31 before tutoring and 4.44 after tutoring. A t-test shows that these values are not significantly different from each other (p= .43). In addition, pre- and post-tutoring motivation levels are very significantly correlated ($R(97) = .69$, $p < .0001$). In this work we use the pre-tutoring motivation scores, and discard the post-tutoring scores.

5.1.2 Evaluating the Near Far Division

In Section 4.4.1 I described the origin of the questions used for pre- and post-tests in this study, and argued that they could be categorized into sets of "near" and "far" transfer questions based on their source. This division is plausible and follows my own previous work (Ward and Litman, 2008); however a formal tagging study to label individual questions as "near" or "far" was beyond the scope of the current study.

Here I provide further support for this categorization by comparing the pre- and post-test correctness percentages for the two sets of questions. This analysis shows that the far transfer questions become relatively more difficult only after tutoring, as would be expected if they were effectively non-isomorphic to the tutored problems.

Table 10 shows percentage correct on the pre- and post-tests for both the 26Q (near) and 18Q (far) question sets. Note that before tutoring, the "far transfer" questions actually have higher mean correctness (e.g. compare .54 near and .65 far). *After* tutoring, however, the "far transfer" questions have lower mean correctness figures (e.g. .76 near vs .70 far).

This relationship is also shown graphically in Figure 5.

	Mean % Cor.	
	Near	**Far**
Pre:	0.542	0.651
Post:	0.764	0.700

Table 10: Percentage correct on 26Q("near") vs 18Q("far") question sets. N=99

Table 11 shows p-values for an Anova predicting percentage correct by test phase (pre-test or post-test), gain type (near transfer or far transfer) and their interaction. This shows a significant effect for test phase, reflecting that students learned between the pre and post tests. There is also a significant interaction between gain type and test phase, suggesting that this gain was significantly different for near and far transfer questions.

A post hoc TukeyHSD analysis shows that correctness on *near* transfer questions is significantly different between the pre- and post-tests ($p < 0.000$). However the difference between pre- and post-test correctness is not significant ($p=0.107$) for *far* transfer questions.

	Test Phase	**Gain Type**	**Test Phase x Gain Type**
Pct. Correct	0.000	0.135	0.000

Table 11: Anova explaining pct. correct by test phase (pre or post), gain type (near or far transfer), and their interaction.

In this work we define "far transfer" questions to be those which are dissimilar to the tutored problems, rather than intrinsically harder. We do not expect them to be harder to solve before tutoring, using previous knowledge. Therefore it seems sensible to assume that these problems may not be relatively more difficult than near transfer problems until after the tutored material has been presented. The fact that our far transfer questions become relatively harder only *after* tutoring lends additional support to our division of questions.

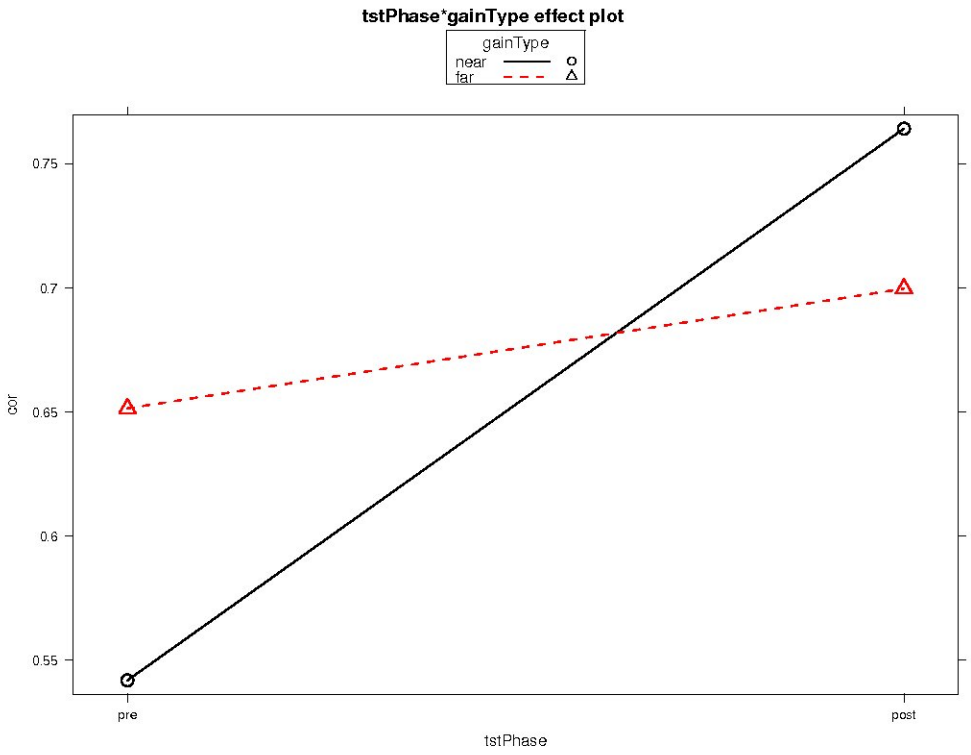

Figure 5: Pre- and post-test correctness for near and far transfer questions, showing that far-transfer questions get relatively harder after tutoring.

5.1.3 Evaluating the Cognitive Load Measures

As described in Section 4.4.2 I developed two metrics intended to measure cognitive load. One measure was based on records of reading time, while the other measure was based on self-reports of reading difficulty. Both the reading speed and reading difficulty measures were converted into cognitive load measures by averaging them per student and normalizing by subtracting a baseline. This produced two measures of cognitive load per student. I will evaluate their reliability by examining correlations between them at the student level.

A Pearsons correlation shows that these two measures of cognitive load are significantly, but not highly, correlated ($R(97) = -0.277$, $p < 0.006$). The correlation is negative because the reading speed measure becomes smaller with higher cognitive load, while the self-report measure becomes larger with higher cognitive load.

Although these two measures are significantly correlated, Cronbach's alpha (run after reversing the scale on the self-report measure) suggests that they have extremely low reliability $\alpha < 0.000$.

The lack of consistency between these two measures of cognitive load suggests that one of them may be less reliable than the other. We assess each measure's reliability separately by calculating its alpha between text blocks for each student. That is, for each text we calculate the cognitive load measure separately on each text block for each student, then calculate alpha across the set of text blocks. We do this separately for the various texts because they contain different numbers of blocks. Results are shown in Table 12.

Test	Target Text	# Students	# Blocks	Alpha
Reading Speed	Again	33	11	.87
Reading Speed	Ref-low	34	6	.71
Reading Speed	Ref-high	32	7	.66
Reading Difficulty	Again	33	11	.80
Reading Difficulty	Ref-low	34	6	.75
Reading Difficulty	Ref-high	32	7	.78

Table 12: Between block alphas by cohesion measure and text

Surprisingly, both measures show high internal reliability. All of the alphas are above the .7 threshold for reliability except for the reading speed measure, calculated on the high

cohesion reflective text.

These results suggest that the measures of cognitive load used to test hypotheses four and five may each be reliable, but address different aspects of cognitive load. This further suggests that the results presented in Sections 5.6.1 and 5.6.2 should be interpreted with caution.

5.2 DIVIDING SUBJECTS BY MOTIVATION

The motivation results are on a continuous scale, but for later analysis we want to divide subjects into three discrete categories by motivation level [1]. I do this first by noticing that the midpoint of the response scale is 3.5. Subjects who feel un-motivated to learn physics will probably choose average scores below this midpoint, so I set the first threshold at 3.5 and call the subjects with motivation scores below this the "low" motivation group. Summary statistics for students who averaged below 3.5 on the motivation scale are shown on the top line of Table 13.

Students who chose above 3.5 are summarized on line two of Table 13. We use the median score of these students to divide them into "middle" motivation subjects who scored below 4.75 (and above 3.5) and "high" motivation subjects who scored above 4.75. These students are summarized on the last two lines of table Table 13. Note that this division is not sensitive to the choice of "mean" or "median." Note also that these thresholds also divide the data set into fairly even thirds.

The distributions of students in each motivation group among pre-test categories (which were described in Section 4.2) and experimental conditions (which were described in Section 4) are shown in Tables 14, 15 and 16.

In Sections 5.3.2, 5.4.2 and 5.5.2 I use these categories to show interactions between motivation, student knowledge and textual cohesion.

[1] Note that subjects were also divided into three knowledge categories according to pretest score, however only the top and bottom pre-test groups were retained for the study.

	Min.	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**	**N**
below 3.5 (low mot)	1.500	2.438	2.750	2.750	3.250	3.500	28
above 3.5	3.750	4.500	4.750	4.930	5.500	6.750	71
Mid Mot	3.750	4.250	4.500	4.382	4.500	4.750	36
High Mot	5.000	5.250	5.500	5.493	5.750	6.750	35

Table 13: Summary statistics for student motivation

preTest cat	again	hiRef	lowRef	tot
low preTesters	2	4	7	13
high preTesters	4	6	5	15
	6	10	12	28

Table 14: Distribution of lowMot subjects

preTest cat	again	hiRef	lowRef	tot
low preTesters	7	3	6	16
high preTesters	4	8	8	20
	11	11	14	36

Table 15: Distribution of midMot subjects

preTest cat	again	hiRef	lowRef	tot
low preTesters	6	6	4	16
high preTesters	10	5	4	19
	16	11	8	35

Table 16: Distribution of highMot subjects

5.3 HYPOTHESIS ONE: ABSTRACTIVE REFLECTION IMPROVES LEARNING

Before examining the hypotheses in detail, it is worth noting that students learned overall as a result of the experiment. Mean correctness (a count of the 44Q questions answered correctly, averaged over all students) rose from 25.8 on the pre-test to 32.5 on the post-test. An Anova predicting correctness by test phase (pre- to post-test) shows that this difference is highly significant ($p < 0.000$). Similarly, mean correctness on the delayed post-test was 31.1. Repeating the Anova shows that students also learned significantly between the pre- and delayed post-test ($p < 0.000$).

Our first hypothesis was that abstractive reflection would improve learning, ie: that there would be an overall advantage to reflection without regard to the cohesiveness of the reflective text.

5.3.1 Initial Results

To test if there was an overall effect for reflection we combined the high and low cohesion conditions into one "ref" condition. We ran an Anova with normalized learning gain as the dependent variable and condition ("ref" vs "noRef") as the independent variable. Table 17 shows results for this Anova, for each of the six measures of normalized learning gain described in Section 4.4.1. The second column shows the p-value for the Anova. The third and fourth columns show mean normalized learning gain for the control and reflection groups, respectively, while the fifth and sixth column show the number of subjects in each group. Note that while the mean learning gains often favor reflection, none of the results approach statistical significance.

Gain Type	pVal	Mean NLG		N	
		noRef	ref	noRef	ref
allGain	0.160	0.314	0.379	33	66
nearGain	0.518	0.487	0.521	33	66
farGain	0.745	0.016	0.049	33	66
delAllGain	0.742	0.310	0.293	33	66
delNearGain	0.885	0.336	0.345	33	66
delFarGain	0.394	0.223	0.142	33	66

Table 17: Anovas explaining NLG by reflection category; all subjects

5.3.2 Motivation Interaction Results

Finding no significant effect for reflection overall, I next tested for interactions with motivation. For this I ran an Anova with normalized learning gain as the dependent variable and both condition ("ref," "noRef") and motivation category ("hiMot," "midMot," "lowMot") as the independent variables. Table 18 shows pValues for these two predictors and their interaction, for all six measures of normalized learning gain. The last column of this table shows a significant interaction between experimental category and motivation for "overall" and "near" learning. Significant p-values are in bold.

subj Group	Gain Type	condCat pVal	motCat pVal	condCat x motCat
all	allGain	0.146	0.303	**0.017**
all	nearGain	0.503	0.513	**0.008**
all	farGain	0.747	0.552	0.576
all	delAllGain	0.745	0.665	0.699
all	delNearGain	0.886	0.950	0.478
all	delFarGain	0.398	0.401	0.835

Table 18: Anovas explaining NLG by reflection and motivation categories; all subjects

Following the method described in Section 4.5, I next divided the corpus into "high" "med" and "low" motivation subjects using the same splits as in Table 18. For each level of motivation and each measure of normalized learning gain, I ran an Anova with NLG as the

independent variable and reflective category ("reflect" vs "no reflect") as the independent variable. Table 19 summarizes the results. For the middle motivation subjects, there is a significant difference in NLG between conditions for overall learning and also for near transfer learning. The mean NLG's shown in columns 4 and 5 show that learning is higher in the reflective group for these measures.

This supports hypothesis one, suggesting that reflection is better than the read again control, but only for our middle motivation subjects.

			NLG			
subj Group	**Gain Type**	**pVal**	**Mean noRef**	**Mean ref**	**N noRef**	**N ref**
lowMot	allGain	0.299	0.243	0.360	6	22
lowMot	nearGain	0.112	0.331	0.521	6	22
lowMot	farGain	0.662	0.077	-0.019	6	22
lowMot	delAllGain	0.753	0.249	0.286	6	22
lowMot	delNearGain	0.231	0.228	0.364	6	22
lowMot	delFarGain	0.491	0.248	0.084	6	22
midMot	allGain	**0.007**	0.185	0.410	11	25
midMot	nearGain	**0.032**	0.378	0.558	11	25
midMot	farGain	0.409	-0.131	0.060	11	25
midMot	delAllGain	0.752	0.264	0.291	11	25
midMot	delNearGain	0.843	0.322	0.344	11	25
midMot	delFarGain	0.929	0.131	0.115	11	25
hiMot	allGain	0.222	0.429	0.359	16	19
hiMot	nearGain	0.076	0.620	0.471	16	19
hiMot	farGain	0.832	0.094	0.114	16	19
hiMot	delAllGain	0.484	0.364	0.302	16	19
hiMot	delNearGain	0.540	0.387	0.325	16	19
hiMot	delFarGain	0.765	0.277	0.246	16	19

Table 19: Anovas explaining NLG by reflection condition, for each motivation category.

5.4 HYPOTHESIS TWO: REFLECTIVE COHESION AFFECTS LEARNING

The second hypothesis was that the cohesiveness of the reflective text would affect how much students learned from it. We expected an overall advantage for high cohesion reflective text relative to low cohesion reflective text.

5.4.1 Initial Results

To answer this question we again perform an Anova, this time explaining NLG by experimental condition, where (in contrast to the test of hypothesis one) the conditions are now "again" "refHigh" and "refLow." This tests the null hypothesis that mean normalized learning gain is equal in all conditions. Results are shown in Table 20.

		NLG					
Gain Type	**pVal**	**Mean Again**	**Mean refHi**	**Mean refLo**	**N noRef**	**N refHi**	**N refLo**
allGain	0.295	0.314	0.397	0.361	33	32	34
nearGain	0.381	0.487	0.559	0.484	33	32	34
farGain	0.947	0.016	0.046	0.052	33	32	34
delAllGain	0.579	0.310	0.323	0.263	33	32	34
delNearGain	0.630	0.336	0.379	0.313	33	32	34
delFarGain	0.479	0.223	0.191	0.097	33	32	34

Table 20: Anovas explaining NLG by Experimental Condition, all subjects

Mean normalized learning gains almost always favor the high cohesion reflective text as predicted by hypothesis two, however none of the differences approach statistical significance.

5.4.2 Motivation Interaction Results

Finding no significant effect for reflective cohesion overall, I again tested for interactions with motivation. I ran an Anova with normalized learning gain as the dependant variable and condition ("again," "refHigh" or "refLow"), motivation ("highMot" or "lowMot") and their interaction as independent variables. Results are shown in Table 21.

subj Group	Gain Type	condCat pVal	motCat pVal	condCat x motCat
all	allGain	0.276	0.339	0.065
all	nearGain	0.360	0.577	**0.038**
all	farGain	0.948	0.550	0.546
all	delAllGain	0.576	0.708	0.186
all	delNearGain	0.631	0.971	0.223
all	delFarGain	0.482	0.440	0.469

Table 21: Anovas explaining NLG by experimental condition and motivation category; all subjects

Table 21 shows that there was a significant interaction between experimental condition and motivation for near transfer learning, and a strong trend for overall learning. This pattern of results is very similar to that shown in Table 18 when explaining NLG by reflective condition.

I next analyze this interaction by dividing the data into motivation categories, and examining the effect of experimental condition on learning separately within each motivation category. I use an Anova predicting normalized learning gain by experimental category, where experimental category is "again" "refHigh" or "refLow."

For middle motivation students, Table 22 shows a significant difference between conditions for both immediate and delayed overall NLG, and a strong trend for near transfer NLG. Next I do a post-hoc TukeyHSD test to determine which of the three conditions were significantly different from each other.

Tables 23 and 24 show results for a post hoc TukeyHSD test on the significant results from Table 22. Given an Anova, the Tukey test does pairwise comparisons of its means and produces adjusted p-values for each comparison. In Table 23 the first column names the comparison being made (eg "refHigh vs again"), and the last column shows the adjusted p-value for that comparison (e.g. 0.019).

For overall gain (Table 23), the high cohesion reflective condition had significantly higher learning gains than the read again control, while the the low cohesion reflective condition

produced a trend toward significance. For delayed overall gain (Table 24), the high cohesion reflective condition is significantly better than the low cohesion condition. Taken together, this suggests that reflective texts can improve learning for "middle" motivated subjects, and that the cohesiveness of the text can affect learning. This supports the second hypothesis that the cohesiveness of the reading affects how well students can learn from it.

These results are for all students, not divided by knowledge level. Next we look in finer detail to see if high and low cohesion texts have different effects if we *do* distinguish by knowledge level.

subj Group	Gain Type	pVal	Mean Again	Mean refHi	Mean refLo	N noRef	N refHi	N refLo
lowMot	allGain	0.583	0.243	0.351	0.367	6	10	12
lowMot	nearGain	0.214	0.331	0.567	0.483	6	10	12
lowMot	farGain	0.443	0.077	-0.152	0.092	6	10	12
lowMot	delAllGain	0.594	0.249	0.227	0.335	6	10	12
lowMot	delNearGain	0.336	0.228	0.314	0.405	6	10	12
lowMot	delFarGain	0.654	0.248	0.009	0.147	6	10	12
midMot	allGain	**0.019**	0.185	0.448	0.379	11	11	14
midMot	nearGain	0.068	0.378	0.605	0.521	11	11	14
midMot	farGain	0.594	-0.131	0.148	-0.009	11	11	14
midMot	delAllGain	**0.043**	0.264	0.417	0.192	11	11	14
midMot	delNearGain	0.136	0.322	0.479	0.238	11	11	14
midMot	delFarGain	0.285	0.131	0.295	-0.027	11	11	14
hiMot	allGain	0.325	0.429	0.388	0.319	16	11	8
hiMot	nearGain	0.162	0.619	0.506	0.423	16	11	8
hiMot	farGain	0.962	0.094	0.124	0.100	16	11	8
hiMot	delAllGain	0.753	0.364	0.317	0.281	16	11	8
hiMot	delNearGain	0.808	0.387	0.338	0.306	16	11	8
hiMot	delFarGain	0.952	0.277	0.252	0.237	16	11	8

Table 22: Anovas explaining NLG by experimental condition, for each motivation category.

	diff	lwr	upr	p adj
refHigh-again	0.263	0.037	0.488	**0.019**
refLow-again	0.194	-0.019	0.407	0.080
refLow-refHigh	-0.069	-0.282	0.144	0.710

Table 23: Post hoc Tukeys allGain, middle motivation

	diff	lwr	upr	p adj
refHigh-again	0.153	-0.0713	0.377	0.229
refLow-again	-0.072	-0.2844	0.139	0.681
refLow-refHigh	-0.225	-0.4375	-0.014	**0.035**

Table 24: Post hoc Tukeys delAllGain, middle motivation

5.5 HYPOTHESIS THREE: REFLECTIVE COHESION INTERACTS WITH KNOWLEDGE

Hypothesis three suggested that the cohesiveness of the reflective reading would affect learning differently for subjects with different knowledge levels. We expected high knowledge readers to learn more from reading a low cohesion reflective text, and low knowledge readers to learn more from reading a high cohesion text.

5.5.1 Initial Results

I examine this question by removing subjects in the "again" condition, and comparing only subjects in the "refHigh" and "refLow" reflective conditions. I first perform an Anova explaining normalized learning gain by pre-test category (high pre-test or low pre-test), experimental category (refHigh or refLow), and their interaction.

	pValues			**Mean Normalized Learning Gain**			
Gain Type	**preTest Cat**	**Exp Cond**	**preTestCat : Exp. Cond**	**hiPre refHi**	**loPre refHi**	**hiPre refLo**	**loPre refLo**
allGain	0.693	0.420	0.458	0.417	0.369	0.352	0.369
nearGain	0.061	0.233	0.286	0.624	0.465	0.505	0.465
farGain	0.228	0.959	0.682	0.013	0.093	-0.030	0.134
delAllGain	0.733	0.279	0.162	0.361	0.268	0.233	0.293
delNearGain	0.474	0.339	0.123	0.439	0.292	0.285	0.341
delFarGain	0.630	0.344	0.445	0.201	0.176	0.028	0.165

Table 25: Anovas explaining NLG by pre-test category and experimental condition; all subjects; expCond = refHi, refLo

Table 25 shows results for this Anova. Each row shows results for a separate Anova, predicting the measure of normalized learning gain shown in the first column. The second and third column show p-values for pre-test category and experimental condition, respectively. The fourth column shows p-values for their interaction. As shown in column four, there are no significant interactions between pre-test category and reflective condition.

5.5.2 Motivation Interaction Results

Finding no significant two way interaction between reflective cohesion and student knowledge level, I next tested for interactions with motivation. For each measure of learning, I ran an Anova explaining NLG by motivation category, pre-test category, reflective condition (high or low cohesion reflective text), and their interactions. Results are summarized in Table 26. Looking at the three-way interaction p-values in the last column, we see that there is a significant interaction for delayed overall learning and delayed near transfer learning.

Gain Type	**motCat**	**preTest**	**Cond**	**motCat :preTest**	**motCat :Cond**	**preTest :Cond**	**motCat :preTest :Cond**
allGain	0.534	0.818	0.365	0.397	0.616	0.663	0.138
nearGain	0.436	0.089	0.186	0.210	0.994	0.466	0.316
farGain	0.584	0.223	0.985	0.583	0.190	0.765	0.376
delAllGain	0.970	0.704	0.260	0.638	**0.032**	0.274	**0.016**
delNearGain	0.886	0.459	0.302	0.902	0.082	0.228	**0.027**
delFarGain	0.449	0.692	0.424	0.699	0.200	0.580	0.355

Table 26: Anova for preTest category, experimental cond, motivation category & interactions. Experimental condition (Cond) = refHi, refLo

Next I examine the two way interaction between knowledge and reflection condition at each of the three levels of motivation. Table 27 shows results for these subdivisions of the data. The left two columns name the motivation group and the type of normalized learning gain used in each Anova. Columns 3, 4 and 5 show the p-values for pre-test category, experimental condition, and their interaction. The last four columns show mean NLG for each of the cells in the Anova. For example "hiPre refHi" means the high pre-testers in the high cohesion reflective condition. Note that the interaction between pre-test category and textual cohesion (in column 3) is significant only for the "middle" motivation subjects. Where there is a significant interaction, the low pre-testers learned more from the low cohesion reflective text ("loPre refLo") than from the high cohesion text ("loPre refHi"). In contrast, the high pre-testers learned more from the high cohesion reflective text.

Figure 6 displays this interaction graphically for delayed overall gain. The dark line shows NLG for high pre-testers, who (as we see in Table 27) had a gain of .50 from the high

cohesion reflective text, and a gain of .10 from low cohesion reflective text. The dashed line shows NLG for low pre-testers, who had a gain of .18 from the high cohesion text, and of .31 from low cohesion text.

For *highly* motivated subjects, on the other hand, there was no interaction between aptitude and textual cohesion: both high and low pre-testers almost always learned more from high cohesion reflective text.

Table 26 shows that there is a significant three way interaction between textual cohesion, domain knowledge and motivation in predicting normalized learning gain. Table 27 explains this three way interaction by showing that there is a significant two way interaction between knowledge and textual cohesion, but only for middle motivation subjects. This interaction tells us that textual cohesion affects NLG differently for subjects in different knowledge categories, but it doesn't tell us between which conditions NLG was significantly different. I address this question using the post-hoc TukeyHSD test.

Table 28 shows post-hoc TukeyHSD results for the interactions shown in Table 27 which were significant or a trend. Column three of Table 28 compares low pre-tester's NLG between the high cohesion and low cohesion reflective text. Column four makes the same comparison for the high pre-testers (other paired comparisons are also made by the TukeyHSD test, but have no theoretical interest and so are not reported). High pre-testers learn significantly more from high cohesion than from low cohesion. Low pre-testers had higher mean NLG under low cohesion for every measure of learning (see Table 27), however these differences are not significant. So, we can say that cohesion affects low knowledge students differently than high knowledge students, and that high cohesion does not benefit low knowledge students as it does high knowledge students. However we cannot say that the low knowledge students learned significantly more from low cohesion in this study.

		pValues			Mean Norm. Learning Gain			
subj Group	**Gain Type**	**preTest Cat**	**Exp Cond**	**preTestCat : Exp. Cond**	**hiPre refHi**	**loPre refHi**	**hiPre refLo**	**loPre refLo**
lowMot	allGain	0.987	0.859	0.586	0.330	0.383	0.394	0.348
lowMot	nearGain	0.840	0.502	0.756	0.554	0.586	0.508	0.464
lowMot	farGain	0.910	0.115	0.707	-0.153	-0.152	0.160	0.043
lowMot	delAllGain	0.676	0.279	0.339	0.180	0.300	0.372	0.309
lowMot	delNearGain	0.800	0.385	0.397	0.289	0.350	0.484	0.349
lowMot	delFarGain	0.374	0.497	0.630	-0.077	0.138	0.116	0.169
midMot	allGain	0.535	0.284	0.071	0.479	0.364	0.306	0.476
midMot	nearGain	0.469	0.247	**0.033**	0.668	0.437	0.484	0.571
midMot	farGain	0.297	0.359	0.309	0.154	0.133	-0.200	0.246
midMot	delAllGain	0.638	**0.006**	**0.003**	0.506	0.180	0.102	0.312
midMot	delNearGain	0.710	**0.021**	**0.006**	0.584	0.200	0.132	0.381
midMot	delFarGain	0.910	0.146	0.255	0.365	0.110	-0.153	0.142
hiMot	allGain	0.158	0.259	0.484	0.419	0.362	0.392	0.246
hiMot	nearGain	**0.035**	0.378	0.980	0.636	0.397	0.544	0.301
hiMot	farGain	0.228	0.907	0.479	-0.0129	0.237	0.073	0.127
hiMot	delAllGain	0.524	0.715	0.895	0.348	0.291	0.324	0.238
hiMot	delNearGain	0.517	0.771	0.968	0.386	0.299	0.344	0.267
hiMot	delFarGain	0.635	0.889	0.842	0.273	0.235	0.281	0.193

Table 27: preTest/expCond interactions for diff motivation groups. expCond = refHi, refLo

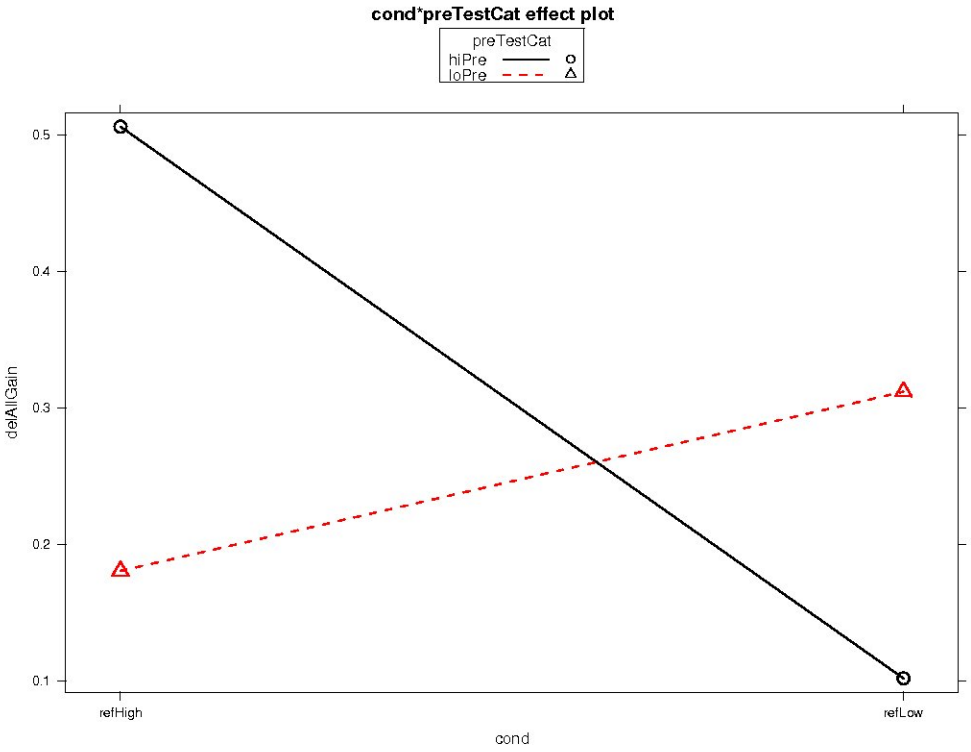

Figure 6: Interaction of cohesion and pre-test category, showing that high pre-testers had higher delayed overall NLG from high than low cohesion, and low pre-testers had higher delayed overall NLG from low than high cohesion.

subj Group	**Gain Type**	**loPre hiRef vs loRef**	**hiPre hiRef vs loRef**
midMot	allGain	0.796	0.222
midMot	nearGain	0.648	0.132
midMot	delAllGain	0.734	**0.001**
midMot	delNearGain	0.705	**0.005**

Table 28: Selected comparisons from post-hoc Tukeys, showing significant NLG difference between cohesion conditions for high pre-test but not low pre-test students.

5.6 SECONDARY HYPOTHESES

5.6.1 Hypothesis Four: Textual Cohesion Affects Learning Through Inference

My fourth hypothesis was that textual cohesion would affect inference during reading. That is, based on work of McNamara's (e.g. (McNamara and Kintsch, 1996)) described in Section 2.4, I expected that if learning was improved by low cohesion text, it would be because the gaps in the text had spurred inference. If this were true, low cohesion text should be accompanied by *increased* cognitive load.

An alternative theory was offered by Kalyuga and Ayres (2003). Kalyuga and Ayres thought of the cohesion-reversal effect as an instance of schema interference. In this view, text with less cohesion requires less reconciliation with existing mental schema, and so imposes less extraneous cognitive load. By this view, the low cohesion text in our study should be accompanied by *decreased* cognitive load.

As shown in Section 5.5.2, our results did suggest that textual cohesion influenced learning, particularly for subjects with middle motivation levels. For those subjects, there was a significant interaction between knowledge and textual cohesion. Students with low pre-test scores had higher mean NLG with low cohesion than high cohesion text, although this difference was not significant. I now address the question of whether this interaction in learning gains is accompanied by a similar interaction in cognitive load during reading, and is therefore explainable by either the inference-making or schema-interference theories described above.

To address this question I included two measures of cognitive load in the study. The first was a measure of reading speed. A decrease in a student's reading speed relative to a baseline indicates increased cognitive load. The second was a self-report measure of reading difficulty. An increase in rated difficulty relative to a baseline indicates increased cognitive load. These measures were described more fully in Section 4.4.2.

Table 29 shows results of an Anova explaining cognitive load by knowledge level (high or low pre-test) and reflective condition (high or low cohesion text). This is a similar setup to the Anova used to test Hypothesis 3, with the exception that the dependant variable is cognitive

	pValues			Cog. Load: Avg. Reading Diff.			
Subj. Group	**preTest Cat**	**Exp. Cond**	**preTest x Exp. Cond**	**hiPre refHi**	**loPre refHi**	**hiPre refLo**	**loPre refLo**
All	0.638	0.266	**0.015**	0.933	0.470	0.410	0.722
Low Mot.	0.641	0.658	0.092	1.044	0.286	0.340	0.738
Mid Mot.	0.247	0.139	0.707	1.017	0.635	0.542	0.372
High Mot.	0.132	0.447	**0.014**	0.665	0.511	0.233	1.217

Table 29: Anova explaining self-reported cognitive load by pre-test and experimental condition (refHi or refLo)

load rather than NLG. Therefore, if there was a significant interaction between knowledge and cohesion that affected learning from text, and if cohesion was affecting inference, then we might also see an interaction affecting cognitive load.

This Anova uses the self-report cognitive load measure, which was calculated for each student as average difficulty in the reflective reading minus average difficulty in the physics pre-reading. The physics pre-reading was chosen as a baseline in preference to the non-physics warmup reading in order to isolate the effects of textual cohesion. There were no significant results when using the reading-speed measure of cognitive load.

Table 29 shows no main effect for either pre-test category (knowledge) or for experimental condition (textual cohesion). However, there are significant interactions between knowledge and cohesion for all students, and also for the highly motivated students (rows one and four in Table 29).

Note that in those student groups which have a significant interaction, cognitive load for HIGH pre-testers is higher when reading the high cohesion text, and lower when reading the low cohesion text. Cognitive load for LOW pre-testers is higher when reading the low cohesion text. Figure 7 shows this interaction for the highly motivated students in Table 29.

This result seems to favor Kalyuga's interpretation of the reverse cohesion effect, how-

ever it should be noted that differences in cognitive load were predicted for the *middle* motivation subjects, who had shown a significant learning interaction between knowledge and cohesion. The results in Table 29 do not show a matching cognitive load interaction for middle motivation subjects.

Hypothesis four is therefore not supported. Although middle motivation students' learning was significantly affected by the cohesiveness of the reflective text, that interaction was not accompanied with a similar interaction in cognitive load, so this cognitive load measure does not allow us to attribute the difference in learning to differences in inference during reading.

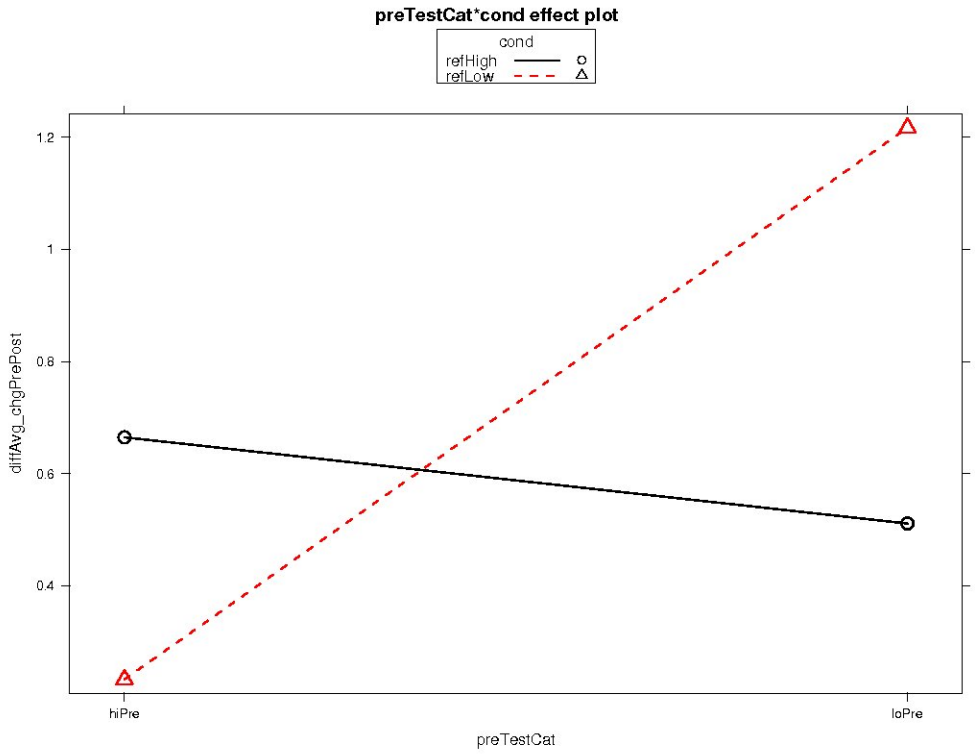

Figure 7: Cognitive Load for High Motivation Subjects Reading High and Low Cohesion Text, showing that high pre-testers had higher load from high cohesion than low cohesion text.

5.6.2 Hypothesis Five: Motivation Affects Inference

Hypothesis five was that motivation levels would affect inference from text. I expected students with higher levels of motivation to engage the text more actively, make more inferences, and so have higher cognitive load.

As in Section 5.6.1, I look for evidence of inference by examining measures of cognitive load. Recall from Section 4.4.2 that I collected reading times for each block of text, which were converted into words-per-minute scores. I calculated a cognitive load measure by subtracting the reading speed for the non-physics warmup reading from the speed on the post-tutoring text. The warmup reading was chosen as a baseline in preference to the introductory physics reading because we are interested in the effects of physics-related motivation. Subtracting a baseline adjusts for individual differences in reading speed, so that the resulting number shows more clearly the effect of the difference in topic motivation. Note that a lower number on this measure means that the student read more slowly relative to the baseline, and so additional cognitive load due to motivation was higher.

Table 30 shows median cognitive load for "low" "middle" and "high" motivation subjects, using the reading speed measure. Note that the low motivation students also have the lowest cognitive load when reading, shown by their higher scores.

Table 31 shows results for an Anova explaining cognitive load by motivation category and type of text: high cohesion or low cohesion. As shown in Table 31 cognitive load showed significant differences by motivation category, but not by type of text. There was also no interaction between motivation category and text type.

	lowMot	**midMot**	**hiMot**
All Text	289.508	71.226	105.043
Low cohesion text	179.137	71.604	129.564
High cohesion text	285.445	59.729	93.191

Table 30: Cog load by motivation Category, showing lower load for lowMot students (higher number = lower load)

Cog Load Measure	motCat	Exp. Cond.	motCat x expCond
wpmMed	**0.002**	0.380	0.140

Table 31: Anova explaining Cog. Load by motivation category and experimental condition (refHi or refLo)

I next repeated the Anova using only motivation category as an independent variable, and performed a post-hoc TukeyHSD test to determine which of the three motivation categories were significantly different. This test showed that cognitive load was significantly higher for high motivation students than for low motivation students ($p = 0.020$). Cognitive load was also significantly higher for middle motivation than low motivation readers ($p = 0.002$).

Tests using the self-reported "reading difficulty" measure of cognitive load failed to show significant differences between conditions. Results for the reading speed measure in Table 31, although weakened by the lack of corroboration from the self report measure, suggest that cognitive load during reading is higher for more highly motivated students. This tends to support hypothesis five, which suggested that highly motivated students would engage the text more actively.

6.0 USING COHESION TO EVALUATE KNOWLEDGE LEVEL

The preceding results have shown that it can be useful to read a reflective text after tutoring in qualitative physics, but that the benefit varies with the knowledge and motivation level of the student. In particular, we have evidence that, for moderately motivated students, a high cohesion text will help students who have a high amount of knowledge to learn. However, a text with low cohesion may be better for students with a low amount of domain knowledge, although that advantage was not significant in the current study. In any actual implementation of a post-tutoring reflective text, therefore, it would be useful to accurately identify student levels of knowledge or motivation in order to personalize the text.

In the current study, I categorized students as having "low" or "high" domain knowledge based on their pre-test score. In a deployed system, however, using pre-test alone might not be optimal because it ignores any learning from tutoring. Inserting another post-test between tutoring and the reflective reading would probably also be sub-optimal, because frequent tests are time consuming and probably onerous for the student. A better solution, if possible, would be to use an estimate of student knowledge that had been updated with information from the tutoring dialog.

Researchers have examined a variety of features that are associated with learning from tutorial dialog. For example, Core, Moore, and Zinn (2003) found that dialog interactivity was correlated with learning in human-human tutorial dialog. Litman and Forbes-Riley (2006a) found that certain types of dialog act predicted learning, and Dzikovska et al. (2008) found that tutorial restatements of correct tutorial response to student answers were also associated with learning from tutoring. In peer learning, Kersey et al. (July, 2009) have found that initiative shifts were associated with learning from dialog.

In my own previous work, I have also developed several methods which are useful in

gauging student learning during natural language tutoring. These methods are interesting in the present context both because they were tested on corpora that were very similar to the current one, and also because they use dialog features which are closely related to the elements of textual cohesion that were manipulated in the current study.

In Section 6.1 I describe these previously developed measures of dialog cohesion, and summarize their results on previous corpora of tutoring dialogs. Then, in Section 6.2, I show that they also produce significant correlations with learning in the current corpus.

6.1 DIALOG COHESION AND LEARNING: METRIC AND PRIOR RESULTS

In (Ward and Litman, 2006) and (Ward and Litman, 2008) I created measures of cohesion in dialog by adapting the concept of a "cohesive tie," that had been used by Halliday and Hasan (1976) to describe cohesion in text. In *Cohesion in English* Halliday and Hasan suggest that the cohesiveness of a text can be measured by counting the number of "cohesive ties" that it contains, where a "cohesive tie" consists of two words joined by a relationship such as exact repetition, repetition of synonyms, repetition of superordinate class terms and so forth.

I measured cohesion in dialog by counting three kinds of cohesive tie between tutor and student: "token," "stem group," and "semantic similarity" repetition, which correspond roughly to the three types of Halliday and Hasan tie mentioned above.

At the "token" level, a cohesive link was counted if exactly the same word form (after stripping punctuation, and ignoring case) appeared in both one turn and the next. Results were collected at this level both with and without stop-words being counted. This corresponds to the first of Halliday and Hasan's reiteration types.

At the "stem" level, a link was counted if two words in consecutive turns were given the same stem by a standard Porter stemmer (Porter, 1997). Table 32 gives examples of how tokens are grouped by stems. Tokens to which the stemmer assigns a common stem appear together in the second column, with their stem in the first column.

Table 33 shows how counting cohesive ties at the token or stem level can affect the

Stem Group	Tokens
packag	package, packages
packet	packet, packets
speed	speed, speeding, speeds
veloc	veloc, velocities, velocity, velocitys
horizont	horizontal, horizontally
displac	displace, displaced, displacement, displacements, displacing
find	find, finding
so	so
thu	thus

Table 32: Examples of how tokens are grouped by stems (Table 1 from Ward & Litman 2006)

amount of cohesion counted [1]. Two consecutive turns are shown at the bottom of the table. The three columns at the top of the table show the matches counted at each level, and their total count. For example, the "Token w/stop" level counts 14 exact word repetitions. The "token, no stop" level matches tokens after removing high frequency "stop words." This level counts 9 cohesive ties between these turns. The "stem, no stop" level matches stems after removing stop words. This level counts 11 cohesive ties: the same 9 as at the "token, no stop" level, plus the additional stems "acceler" and "vertic." This allows the stem level to match the tokens "accelerates" to "acceleration," and "vertical" to "vertically." These matches were not found at the token levels.

At the "semantic similarity" level I counted a cohesive tie whenever one utterance and the next had different words with similar meanings, with similarity measured using WordNet (Miller et al., 1990). WordNet is a large semantic lexicon with several useful features. First, it groups words into groups of synonyms called "synsets." Second, it organizes synsets into an "is-a" (hypernym/hyponym) taxonomy. If we know the sense in which a word is being

[1] The algorithm and results presented in this section were originally published in (Ward and Litman, 2006) and (Ward and Litman, 2008). Cohesion examples in this section were taken from previous Itspoke tutoring dialogs which included essay submissions. Results on the current corpus which has no essays are in Section 6.2.

Token w/stop(14)	**Token, no stop (9)**	**Stem, no stop(11)**
packet, horizontal, the, it, is, of, only, force, acting, on, there, will, still, after	packet, horizontal, only, force, acting, there, will, still, after	packet, horizont, onli, forc, act, acceler, vertic, there, will, still, after

Student Essay	No. The airplane and the packet have the same horizontal velocity. When the packet is dropped, the only force acting on it is g, and the net force is zero. The packet accelerates vertically down, but does not accelerate horizontally. The packet keeps moving at the same velocity while it is falling as it had when it was on the airplane. There will be displacement because the packet still moves horizontally after it is dropped. The packet will keep moving past the center of the swimming pool because of its horizontal velocity.
Computer Tutor	Uh huh. There is more still that your essay should cover. Maybe this will help you remember some of the details need in the explanation. After the packet is released, the only force acting on it is gravitational force, which acts in the vertical direction. What is the magnitude of the acceleration of the packet in the horizontal direction?

Table 33: Two consecutive turns, counting cohesive ties at the token and stem levels (Table 2 from Ward & Litman 2006)

used, and therefore its relevant synset, we can use WordNet to find its relationship to other synsets in the taxonomy.

Word sense disambiguation was done in a simple way: for each potential pair of words, I chose the senses in which the words were most similar.

I measured semantic similarity as a function of the distance between two concepts in the WordNet hierarchy. In this work, I used the simplest method of measuring this distance: Path Distance Similarity, as implemented in NLTK (Loper and Bird, 2002). This measure calculates the similarity between two concepts as 1/1+N, where N is the number of edges in the shortest path between them. Scores range from zero to one, where zero means no similarity and one indicates exact synonyms. For example, in WordNet the shortest path between "man" and "person" has two edges, and so the similarity of those terms is 1/1+2, or.333.

Finding semantic similarity ties is slightly more complicated than finding lexical reiteration ties because a word in one utterance may have non-zero similarities to several words

Sim	Start		Step 1		Step 2	
	Tok A	**Tok B**	**Tok A**	**Tok B**	**Tok A**	**Tok B**
0.33	man	person	**man**	**person**	man	person
0.13	release	person	release			
0.13	release	elevator	release	elevator	**release**	**elevator**
0.11	velocity	person	velocity		velocity	
0.09	man	elevator		elevator		
0.07	velocity	elevator	velocity	elevator	velocity	

Table 34: Finding the best semantic ties (Table 3 from Ward & Litman 2008)

in the other utterance. Therefore, I created the following algorithm to find the best set of ties. For each word in utterance B, I looked up the WordNet path similarity values between it and each word in utterance A. After collecting the set of all possible word pairs this way, I sorted them by their similarity values. Starting at the high end of the list, for each pair I removed all lower pairs which had the same token in the same position. This process is illustrated in Table 34. To keep the example small, I have selected only two tutor and three student words from the example shown in Table 36. This produces six possible pairs, which are shown in columns two and three of Table 34, sorted by their similarity values

Starting at the top of the list, the algorithm considers first the pair: "man-person." It removes all instances below of "man" in position A and of "person" in position B. This step is shown under "Step 1" in Table 34. In step 2, it moves down to the next remaining pair, "release-elevator." It removes all instances below that of "release" in position A and of "elevator" in position B. There are no pairs remaining to be considered in this example, so it stops and counts two semantic cohesive ties: "man-person" with a similarity of .33, and "release-elevator" with a similarity of .13.

This method can count cohesive ties with a broad range of similarity scores. I investigated whether the stronger ties were more useful by instituting a threshold, and only counting cohesive ties for pairs with similarity values above the threshold. In the example shown in Table 34, a threshold of .3 would count the tie between "person" and "man" but not between "elevator" and "release."

Threshold		
> 0.5	**0.3**	**0**
5-five	motion-contact	remains-same
remain-stay	man-person	man-runner
speed-velocity	decrease-acceleration	force-magnitude
conclude-reason	acceleration-change	summarize-point
package-packet	travel-flying	submit-pull

Table 35: Example Semantic ties (Table 4 from Ward & Litman 2008)

A threshold > .5 counts cohesive ties only for word pairings which are listed in WordNet as being exact synonyms, and which therefore have a similarity score of one (note from the path similarity formula that scores between .5 and 1 are impossible). A threshold reduced to .3 allows cohesive ties with slightly more semantic distance in the pair, and a threshold of 0 allows all pairs found by our algorithm. Examples of cohesive ties counted at each of these thresholds are shown in Table 35. In these examples we can see that the matches counted become more distant and less sensible as the threshold is reduced.

I counted the total number of cohesive ties for each dialog as described above. I then line normalized the count, dividing it by the total number of lines in the dialog. I did this to remove the possibility that the count of cohesive ties correlates with learning simply because the longer dialogs had more cohesive ties. However, neither the total number of tutor turns, student turns, tutor words, or student words are correlated with learning in spoken dialogs with this prior version of our computer tutor (Litman et al., 2006). In the example shown in Table 36, the algorithm counts a total of 10 cohesive ties at the token and stem levels, line-normalizing the count (as if this were an entire dialog) would give a score of $10/2 = 5$.

Finally, I summed the line normalized counts over all dialogs for each student, resulting in a per-student cohesion measure which I correlated with learning.

These metrics were then applied to two corpora of tutoring dialogs that had been collected using the Itspoke tutor in 2003 and 2005. In each corpus I measured the partial correlation between the cohesion measures and post-test score, controlling for pre-test score. Results

Speaker	**Utterance**
Student	Before the **release** of the **keys**, the man's and the keys velocity are the same. **After** the release the **only force** on the keys and **man** is downward force of earth's **gravity**, **so** they are in freefall. We **can** ignore the **forces** that the air exerts on these objects since they are dense. Therefore, at every **point** in time the keys will remain in front of the man's face during their whole trip down.
Tutor	**So** you **can** compare it to your response, here's my summary of a missing **point**: **After** the **release**, the **only force** on the **person**, **keys**, and elevator is the **force** of **gravity**. Kindly correct your essay. If you're finished, press the submit button.
Level	**Cohesive Ties Counted between Utterances, at each level**
Token	so-so, release-release, point-point, only-only, keys-keys, gravity-gravity, can-can, after-after, force-force
Stem	forces-force
Sem	man-person

Table 36: Token, Stem, and Semantic Similarity Sem Matches (Table 2 from Ward & Litman 2008)

Corpus	preTest Cat.	Gain type	Threshold	p-value	Cor.
2003	High	overall	0.30	**0.038**	0.899
2003	High	overall	0.00	**0.002**	0.894
2003	Low	overall	0.30	**0.004**	0.689
2005	Low	overall	0.30	**0.011**	0.613
2005	Low	far	0.00	**0.039**	0.519

Table 37: Partial correlations between learning and dialog cohesion. Low pre-test students (extracted from Table 5 in Ward & Litman 2008).

were slightly different between the 2003 and 2005 corpora. In the 2003 corpus the measure of lexical cohesion was significantly correlated with overall learning for low (below mean) pre-testers. The measure of semantic similarity cohesion was also significantly correlated with overall learning for low pre-testers, and in addition correlated with learning for high pre-testers, with the correlations becoming stronger when using lower similarity thresholds.

In the 2005 corpus, the measure of lexical cohesion correlated with overall learning for low pre-testers, just as it had in the 2003 corpus. The semantic similarity cohesion measure also correlated with overall learning, and in addition correlated with far transfer learning at low thresholds (the near-far transfer distinction had not been available in the 2003 corpus). A portion of the semantic similarity results is reproduced in Table 37, for comparison to results on the current corpus, which are described in the next section.

6.2 COHESION AND LEARNING IN THE REFLECTION CORPUS

In this section I perform a similar analysis in the new corpus of Itspoke dialogs which was collected for the current study. I use the same measures of normalized learning gain described in Section 4.4.1[2], and report correlations between NLG and dialog cohesion. For consistency

[2]The previous work described in Section 4.4.1 reported partial correlations of dialog cohesion and post-test score, controlling for pre-test score. In this section I report correlations between dialog cohesion and normalized learning gain, for consistency with the analysis in Section 5.

with the previous work in dialog cohesion just described, I report results using a high/low pre-test split. For consistency with the analysis of *textual* cohesion in this thesis, I also report results using the low/mid/high motivation split.

Table 38 shows results for high and low knowledge students, using extreme groups design, as described in Section 4.2. Note that, similar to the results from the 2003 corpus shown in the top two rows of Table 37, the significant results in the 2010 corpus are for high knowledge students. In addition, significance was achieved at the lower semantic thresholds (higher thresholds produced no significant results or trends, and are not shown). Finally, note that the 2010 results were for delayed far transfer learning gains. The near/far division was not available in the 2003 question set, however the 2005 corpus included results for far transfer learning (for low knowledge students). Together, these similarities suggest that this measure of dialog cohesion correlates with learning in the new corpus in a way similar to previous corpora.

Corpus	**preTest Cat.**	**Gain type**	**Threshold**	**p-value**	**Cor.**
2010	Low	allGain	0.10	0.909	-0.017
2010	Low	nearGain	0.10	0.415	-0.125
2010	Low	delFarGain	0.10	0.680	0.063
2010	Low	allGain	0.20	0.824	0.034
2010	Low	nearGain	0.20	0.553	-0.091
2010	Low	delFarGain	0.20	0.578	0.085
2010	Low	allGain	0.30	0.763	0.046
2010	Low	nearGain	0.30	0.679	-0.064
2010	Low	delFarGain	0.30	0.635	0.073
2010	High	allGain	0.10	0.234	-0.165
2010	High	nearGain	0.10	0.960	0.007
2010	High	delFarGain	0.10	**0.033**	0.291
2010	High	allGain	0.20	0.276	-0.151
2010	High	nearGain	0.20	0.880	0.021
2010	High	delFarGain	0.20	**0.067**	0.251
2010	High	allGain	0.30	0.128	-0.210
2010	High	nearGain	0.30	0.877	0.021
2010	High	delFarGain	0.30	0.234	0.165

Table 38: Correlations between learning and dialog cohesion, new 2010 corpus

Table 39 shows results for the same correlations between NLG and dialog cohesion, but

divides students by motivation level. We again see significant results for delayed far transfer learning, at thresholds near 0.3. However, Table 39 shows that these results are for highly motivated students.

Corpus	Motivation Group	Gain type	Threshold	p-value	Cor.
2010	Low	allGain	0.10	0.973	0.007
2010	Low	nearGain	0.10	0.454	-0.147
2010	Low	delFarGain	0.10	0.153	0.278
2010	Low	allGain	0.20	0.750	0.063
2010	Low	nearGain	0.20	0.756	-0.061
2010	Low	delFarGain	0.20	0.115	0.305
2010	Low	allGain	0.30	0.911	-0.022
2010	Low	nearGain	0.30	0.677	-0.082
2010	Low	delFarGain	0.30	0.271	0.215
2010	Mid	allGain	0.10	0.616	-0.087
2010	Mid	nearGain	0.10	0.848	-0.033
2010	Mid	delFarGain	0.10	0.682	0.071
2010	Mid	allGain	0.20	0.609	-0.088
2010	Mid	nearGain	0.20	0.821	-0.039
2010	Mid	delFarGain	0.20	0.765	0.052
2010	Mid	allGain	0.30	0.467	-0.125
2010	Mid	nearGain	0.30	0.987	0.003
2010	Mid	delFarGain	0.30	0.998	0.000
2010	High	allGain	0.10	0.569	-0.099
2010	High	nearGain	0.10	0.408	-0.144
2010	High	delFarGain	0.10	0.059	0.323
2010	High	allGain	0.20	0.810	-0.042
2010	High	nearGain	0.20	0.664	-0.076
2010	High	delFarGain	0.20	**0.042**	0.346
2010	High	allGain	0.30	0.956	-0.010
2010	High	nearGain	0.30	0.800	-0.044
2010	High	delFarGain	0.30	**0.039**	0.351

Table 39: Cohesion-Learning correlations by motivation category

It is encouraging that this measure of cohesion has now produced correlations with learning in three corpora which were collected under different experimental setups over a course of seven years. It suggests that measuring dialog cohesion during tutoring might be a useful addition to a student model that estimated student knowledge at the end of tutoring. This student model could then be used to personalize a post-tutoring adaptive text. It

should be noted, however, that based on evidence from Chapter 5, we would prefer to be able to judge the learning for middle (rather than high) motivation students. This is because those were the students who showed an interaction with textual cohesion, and for whom the benefit of personalized cohesion might be largest. The idea of a post-tutoring adaptive text is expanded in Section 8.3

7.0 RELATED WORK

The study described above drew heavily on previous work in the areas of reflection, text processing and computational linguistics. In this chapter I review relevant work in each of those areas, and show how the current study is related to them.

First I consider previous work in reflection. I review a number of previous studies, and suggest that my work extends them both in domain and in the specific implementation of reflection used. Second, I review related work in text processing, which collectively shows that motivation, knowledge and textual cohesion can all impact what is learned from text. I compare those results to my findings for those three factors. Finally, I review work from the natural language processing literature concerning cohesion in text and dialog, and compare my work in dialog cohesion.

7.1 RELATED WORK IN REFLECTION IN INTELLIGENT TUTORING SYSTEMS

In this section I look at related work in reflection. Section 7.1.1 describes several studies of reflection during problem solving. Section 7.1.2 adds more detail to our previous discussion of the work of Katz and colleagues in reflection following problem solving in Andes, which is the work most closely related to the current study. Following that, in Section 7.1.3 I discuss a broader range of other reflection studies, including those in non-physics domains and with other reflective interventions. Finally, I discuss the current work in light of those related studies.

7.1.1 Reflection In Action

Sections 2.3.2 and 2.3.1 described "reflection-in-action" and "reflection-on-action," and discussed the differences between them. The thesis work described here was of the "reflection-on-action" type. However, several studies have examined "reflection-in-action," usually by means of having students self-explain their own problem solving steps.

For example, the Cognitive Algebra tutor was modified (Aleven and Koedinger, 2002) to require students to select self-explanations for each of their problem solving steps. Requiring self-explanation led to better performance on transfer questions. The authors attribute this in part to improved integration of visual and verbal declarative knowledge.

Similarly, the ALPS learning environment (Corbett et al., 2006) required students to generate explanations for the meaning of each term in an algebra equation. Generating these explanations led to better transfer than selecting them from a menu.

Finally Atkinson et al. (2003) prompted students to identify the principle underlying each step in a worked-out example probability problem, which improved both near- and far-transfer performance.

7.1.2 Reflection In Quantitative Physics

As described in Section 3.1 the current study was carried out in the domain of qualitative physics. Other, closely related work by Sandra Katz and colleagues (Katz et al., 2003, 2007; Connelly and Katz, 2009) has investigated reflection after tutoring in *quantitative* physics.

The first study in this series (Katz et al., 2003) has already been described in Section 2.3.1 to motivate the ideas that reflection could help learning in physics, and that a reflective text could be just as effective as an interactive dialog. In this section, I largely point out a few interesting features of the remaining two.

These studies had several features in common. Reflection questions were given after each of a series of quantitative physics problems in Andes. Typically the reflection question would ask the student to think what would happen to the answer if some feature of the preceding problem were changed. For example, after analyzing the forces involved while pulling a suitcase on frictionless wheels, the reflection question might ask how the forces

would change if there were no wheels. In these studies, qualitative post-test questions were considered far transfer, because tutoring in Andes is largely quantitative. The immediate benefits of reflection were measured using largely qualitative pre- and post-tests administered before and after the relevant unit had been covered in physics class, roughly three weeks apart. Longer term retention was measured using scores on the course final exam, a delayed quantitative problem solving test.

In (Katz et al., 2007), Katz, Connelly and Wilson investigated implementing reflection questions following problem solving in Andes, as described above. Students would enter a response to the question, then either get canned-text feedback or one of two varieties of interactive dialog feedback. Low student participation prevented detailed comparison of the dialog conditions. However, it was determined that students in the reflection condition learned more than students in the non-reflective control when measured by the (near transfer) pre- and post-tests.

Retention and far-transfer were gauged using performance on the delayed final exam, as mentioned above. On this test, students who received canned text feedback actually did marginally better ($p = .07$) than those in the interactive reflection conditions.

Student participation was higher in a second experiment in the same study. In this experiment a new implementation of dialog feedback to reflective questions was compared to standard Andes, without reflective questions. Again, students in the reflection condition learned significantly more, as measured by the near-transfer pre- and post-tests, than students using only Andes. Also similarly to the first experiment, there was no difference in far-transfer learning gains between conditions.

In (Connelly and Katz, 2009), Connelly and Katz again improved the reflective dialog feedback. In an attempt to increase transfer to quantitative problems, the new extended feedback dialogs asked both qualitative and quantitative questions. In addition, some new "what-if" reflection questions were designed to explicitly compare the Andes problems to other problem scenarios. As in the first experiment, unproductive student behavior limited the types of analysis that could be done. However, significant results were obtained by regressing outcome measures (learning gain or final exam score) against number of reflection dialogs completed and other independent measures.

Near transfer learning gains, as measured by the largely qualitative pre- and post-tests, were significantly predicted by reflective dialog completion. In addition, dialog completion was a significant predictor of learning gains on the subset of quantitative questions.

Far transfer learning, as before, was measured by scores on the quantitative final exams. The number of reflective dialogs completed was a positive and significant predictor of final exam score. In both the near- and far- transfer results, the significance of KCD completion disappeared if QPA was added to the regression. These results were the first indication of a longer term benefit to reflection in this series of studies.

Note that in the two studies just described, the benefit of reflection seemed to apply to all students. Statistical results were obtained without subdividing students by pre-test score or motivation.

In (Katz et al., 2003), Katz, Allbritton and Connelly found both that students who reflected learned more than those who didn't (Experiment 2) and also that among students who received reflective dialog, students who participated in more dialogs, or whose dialogs had more of certain features such as abstraction from the previously tutored problem, learned more (Experiment 1). The benefit of reflection in these experiments was shown for the group of all students.

7.1.3 Reflection In Other Domains

A wide variety of interventions have been proposed to encourage student reflection in various contexts. For example, systems have been built that encourage learners to interact with a computer tutor's student model(Cimolino et al., 2003), or as "learning companions" that encourage reflection following interaction with a tutor (Goodman et al., 1998). Interventions have also taken the form of scripts for reflective dialogs among students following group work (Frederiksen and White, 1997), or of software tools to scaffold reflective inquiry in classroom learning (Kyza et al., 2002).

In this section I review a subset of these studies that have been evaluated either in terms of learning gains or of reflective behavior, and compare them to my current work. I categorize these studies along two dimensions, first according to the visibility of the reflection produced

by the student, and second according to the type of guidance given to the student. Taken together these studies will suggest that explicit reflection is better than implicit reflection, and that highly personalized guidance may not be better than more generalized guidance.

The first division is between "explicit" and "implicit" reflection. In an explicit reflection condition, the student is required to reflect out loud or on paper, while in an implicit reflection condition the student does not produce any evidence of reflection, and may or may not be reflecting at all. For example in tutorial dialog, when the tutor points out a mistake, that is an opportunity for the student to reflect on the thought processes that arrived at the error (Tchetagni et al., 2007). If the tutor asks the student to do this out loud (or on paper) the reflection is explicit. Otherwise, the reflection (if it happens at all) is implicit.

The second division is between different amounts of specificity in the guidance provided for reflection. I divide this dimension into three levels. The most specific guidance is "personalized" to match the student's particular errors or misconceptions. An example of this might be if a dialog-based tutor were to ask, following an incorrect response, how the student had arrived at that answer. I term the next category along this dimension "guided" reflection. In this category the tutor asks the student to reflect about some general topic or issue, but the selection of topic is not chosen to benefit that specific student. The final category along this dimension is called "unguided." Here the student is not given a specific topic for reflection, but must choose one independently.

These categories are shown as the horizontal and vertical axes in Table 40. Next I will describe several studies which fit in the various cells of Table 40, and the results of their evaluations. The studies discussed differ on many dimensions, and would be categorized differently in a system which accommodated their many other important differences. This system (shown in Table 40) was chosen to highlight a potentially interesting trend among the patterns of significance in these studies.

Tchetagni et al. (2007) describe a modification to their Prolog-tutor that encourages explicit reflection. In its standard form, Prolog-tutor engages the student in a tutorial dialog concerning problem solving in Prolog. For example, the tutor might ask the student which element of a knowledge base should be included in a Prolog statement. The student might respond "I don't know," and be given a remedial sub-dialog. This form of Prolog-tutor

points out individual student errors at a fine level of detail, but does not ask the student to explicitly reflect on them (although they may do so covertly). It is shown as "Prolog-tutor" in Cell 2 of Table 40.

In its modified form, Prolog tutor asked tutorial questions like above, but also asked a variety of reflection questions. For example, students might be asked to justify their previous response, or to explain if they agreed or disagreed with a tutor statement. Students were asked to respond aloud to these questions in "think aloud" format, and their answers were recorded for analysis. Because this version asks students to reflect explicitly in response to specific prompts, is is shown in Cell 1 of Table 40 as "Prolog-tutor-ER."

An analysis of student responses suggested that Prolog-tutor-ER caused reflective thinking, however the authors noted some qualifications to this result. For example, Prolog-tutor-ER did not seem to trigger much confusion among students, as had been expected. Also, the authors note that the questions were so specific that students seemed unable to see their higher goal, and often didn't seem to realize that they were doing "reflection" at all. This suggests that the effectiveness of reflection may be compromised by making the reflective guidance too specific and personalized.

Gama (2004a) investigated adding a Reflection Assistant (RA) to "MIRA," an algebra word problem solving environment. The focus of reflection in this study was the students' knowledge monitoring accuracy (KMA). Before each problem in MIRA, students would estimate various aspects of their knowledge and ability to solve the upcoming problem. After each problem, the reflective assistant would compare their performance to their predictions, and rate their accuracy. This combination, with a directed (but not personalized) reflection target and explicit reflection, is shown as "MIRA-RA" in Cell 3 of Table 40. In the control condition, students solved algebra problems in MIRA without interacting with the reflection assistant. Mira gave feedback at the answer level rather than at the problem-solving step level, then allowed students to see a correct solution after problem solving (Gama, 2004b). This condition offered an opportunity for implicit, guided reflection, and is shown as "MIRA" in Cell 4 of Table 40.

Results suggested that students in the reflection condition performed significantly better on several measures of problem solving performance (Gama, 2004a). Pre- and post-tests were

not given, however students in the reflective condition got a greater percentage of problems correct and "almost correct" while working in MIRA. A measure of KMA also improved in the reflective condition, but not significantly.

Lin and Lehman (1999) gave students reflection prompts while they solved problems in experimental variable control. The students set up biology lab experiments with a simulated biologist named Paula. The reflection prompts asked students to provide reasons for their actions (*reason* prompts), to explain the rules or procedures they used (*rule based* prompts), or to reflect on their feelings (*emotion focused* prompts). These prompts guided the student to reflect on certain topics, but were not personalized to address the student's knowledge level, so this condition is shown in Cell 3 of Table 40 as Paula-ref.

The control group solved problems without receiving prompts, and received feedback after all the problems were completed. This condition is shown as "Paula" in Cell 4 of Table 40.

The evaluation in the Lin and Lehman (1999) study involved measuring performance on simulated test problems after tutoring, where some problems were contextually similar to the tutored problems ("near" transfer) and some problems were contextually dis-similar to the tutored problems. Both conditions did equally well on the near transfer problems, but the reflection condition that received "reason" prompts did significantly better than the other groups on far transfer problems.

Lee and Hutchison (1998) presented reflection questions to students who had read elaborated or un-elaborated versions of case studies. The cases described how experts balanced chemical equations, and the reflection questions asked students to think about what actions the expert had performed (*case reflection* questions), or why the expert had performed some particular action (*strategy* questions). This condition is shown as "Lee-ref" in Cell 3 of Table 40. In the control conditions, students read the cases without reflection questions. This is shown as "Lee" in cell 4 of Table 40.

Across several experiments, Lee and Hutchison (1998) found that reflection questions produced more learning than the no-reflection control, but that this effect held only for low knowledge students. Students were judged low knowledge if they scored 0 or 1 correct on the pre-test, because both scores were possible without any chemistry knowledge at all. They

hypothesize that "participants who have more pre-knowledge can ask their own questions and therefore do not need the reflection questions as much as those participants without pre-knowledge." Note that this explanation is consistent with the related work described in Section 7, which argues that factors such as increased domain knowledge or increased motivation can lead to higher levels of textual engagement while reading.

Davis (2003) compared *directed* with *un-directed* reflection prompts that were given while students evaluated scientific evidence. Students were required to read an article about heat flow in a fictitious pseudo-scientific "tabloid." They would then critique the claims made and evidence used, and write an analysis in the form of a letter to the editor. Students worked in pairs using the "Knowledge Integration Environment" (KIE), which used a sequence of activity screens to support step-by-step problem solving, and also presented the reflection prompts at various steps. Performance was measured by evaluating the final letters for the coherence of understanding shown in them. Also, quality of reflection was evaluated by examining the student's typed responses to reflection prompts.

Generic prompts were of the form "Our thoughts now are..." they asked the student to reflect about the current activity, but did not dictate a specific topic. Students would complete the prompts by typing responses into a text box in the KIE, so their reflection is "explicit" according to Table 40. The unguided generic prompt condition is shown as "KIE-generic" in Cell 5 of Table 40. An example of a directed prompt is "Claims in the article we didn't understand very well included..." This condition is shown as "KIE-directed" in Cell 3 of Table 40.

Results suggested that students developed significantly more coherent understandings of heat flow from generic rather than directed prompts. This effect was especially strong for students who had scored highly on a measure of "autonomy," indicating that they took responsibility for their own science learning. Students in the directed prompt condition tended to generate less productive reflections, which were in turn correlated with less coherent understanding as shown in the final letter. Davis speculates that because the prompts were dumb, ie. not informed by any type of student model, they were often inappropriate and unrelated to the student's actual difficulties. Generic prompts on the other hand allowed students to expand their repertoire of ideas and identify weaknesses in their own understand-

ing. This effect was especially strong for autonomous students who more readily engaged in reflection.

The thesis experiment reported in the current work required only implicit reflection from students; they were not asked, for example, to write down their own comparisons of the preceding physics problems. In the experimental conditions, students read a reflective reading which directed their attention to certain comparisons between problems, but did not tailor the comparisons to each student. This is shown as "Itspoke-ref" in Cell 4 of Table 40. The control condition, which allowed students to re-read the introductory text in light of the recently tutored problems, is shown as "Itspoke-again" in Cell 6 of Table 40.

In this work, reflection increased learning, but only for students with a certain middle level of motivation. Students with higher levels of motivation were not helped by reading a reflective text. Section 7 reviews related work in motivation and text comprehension which suggests that these highly motivated students may have engaged the control text more actively and so had no need of the additional reflective scaffolding provided in the experimental texts.

Tutor Guidance	**Reflective Response**	
	Explicit	**Implicit**
Personalized	Prolog-tutor-ER (a) Andes-ref-dis (b)	Prolog-tutor (a) Andes (b)
Guided	Andes-ref-noDis (b) MIRA-RA (c) Paula-ref (d) Lee-ref (e) KIE-directed (f)	MIRA (c) Paula (d) Lee (e) Itspoke-ref
Unguided	KIE-generic (f)	Itspoke-again
Reference key: a: (Tchetagni et al., 2007), b: (Katz et al., 2003) c: (Gama, 2004a); d: (Lin and Lehman, 1999) e: (Lee and Hutchison, 1998); f: (Davis, 2003)		

Table 40: Comparing Studies by Type of Reflection

Katz et al. (2003) compared three versions of the Andes homework helper. In the control condition, the students solved quantitative physics problems in the standard version of Andes

and received error feedback when their solutions went awry. This condition is shown in Cell 2 of Table 40 because students were not prompted to reflect, but received personalized correctness feedback that may have led to covert reflection.

In the first experimental condition, students were given a reflection question after tutoring with Andes and asked to discuss their answer with a human tutor. The questions were usually modified versions of problems solved in the preceding Andes session. Some features would be changed in the modified version, and the student asked to consider how the changes would impact the solution. While the reflection questions were not individualized in this condition, the following feedback discussion was, so this condition is shown as "Andes-ref-dis." in Cell 1 of Table 40.

In the second experimental condition, students were given a reflection question after tutoring with Andes, but not allowed to discuss their answer. Instead, they were given a "canned text" response which explained the correct answer. In this condition neither the questions nor the response were individualized for the student, so it is shown in Cell 3 of Table 40 as "Andes-ref-noDis".

Results showed that students in the reflective conditions learned significantly more than those in the control when measured using qualitative questions. The difference was not significant when measured with quantitative questions, and there was no significant difference in learning between the two reflective conditions.

The other two studies of Katz and her colleagues, described in Section 7.1.2, also support the pattern of results seen above. (Katz et al., 2007) showed an advantage for reflection over no-reflection, and the positive correlations found between reflective dialog completion and learning in (Connelly and Katz, 2009) also support an advantage for reflection.

It is now possible to suggest a pattern of results in Table 40. Comparing the two columns we see an overall advantage for "explicit" over "implicit" reflection. The Andes, Mira, Paula and Lee tutors all found an advantage for explicit reflection on their respective evaluation measures. The only exception to this pattern may be the equivocal results for the Prolog tutor, which were attributed to the excessive specificity of the prompts.

Looking at the three rows of Table 40, we see several hints that less guided explicit reflection may be preferable to more individualized explicit reflection. One hint is that

there are no clear examples of effective personalized reflection prompts. As suggested above, reflection prompts in the Prolog tutor may have been too specific to allow the student to step back from problem solving and engage in reflection. In addition, the Andes tutor with human guided reflection caused more learning than the no-reflection control, but seemed no more effective than the less personalized "canned text" feedback condition.

Only one experiment in this sample compared explicit reflection at the "guided" and "unguided" levels. Those results, shown as KIE in Cells 3 and 5 of Table 40 found an advantage for less specific guidance in reflection, particularly for students who had sufficient knowledge to reflect effectively without help.

The current Itspoke study compared implicit reflection at the "guided" and "unguided" levels, and found an advantage for "guided," but only for students with middle motivation. Subjects with higher motivation did not benefit from the reflective text, perhaps because they were prone to reflect effectively without prompting.

Although it is difficult to make confident comparisons across studies that are dissimilar in their outcome measures and implementations of reflection, two broad lessons seem appropriate. First, students will learn more from reflection if required to reflect explicitly. Second, more general instructions to reflect may often be more effective than more specific personalized ones, particularly for students who, for reasons of sufficient knowledge or motivation, may be reflecting by themselves.

7.1.4 Relation To Current Work

The intervention described in the series of studies done by Katz and her colleagues is in many ways similar to the one used in my thesis work. By asking students what would happen if a problem feature were changed, the reflection questions were asking students to compare two different physics questions. Making this comparison could help students differentiate between unimportant surface level features and important deep features of these problems, in the same way that was hypothesized for my intervention. However, there are also some interesting differences. First, the thesis intervention explicitly compared the application of abstract physics laws between different problems, while the feedback to Katz's reflection problems

seems to have largely emphasized the application of laws to the newly modified problem. In addition, in the thesis intervention, the reflective text came *after* all the tutored problems, and made pairwise comparisons between them in various combinations. In contrast, in the Andes reflection intervention, reflection was done *between* the tutored problems, as each problem was completed. Doing reflection after each problem may reduce working memory demands, however it may also reduce the chances that broad comparisons are made between the full set of problem instances.

Another interesting difference between the studies is that reflection after Itspoke only seemed to benefit students with a middle amount of motivation. However, as mentioned above, reflection after Andes benefited all students.

There are many differences between the Itspoke and Andes studies which make it difficult to compare them: they were done in different years, with different students and different measures of learning. It seems likely, however, that a major source of the difference in outcomes between these experiments may be that the two tutors emphasize different aspects of physics. The Andes tutor is a problem-solving homework helper. It provides the student with means to draw free-body diagrams, to define variables and enter equations, and it follows along step-by-step as the student solves quantitative physics problems. Conceptual "on-demand" help is available, and conceptual help was also given by the human tutors during tutoring in (Katz et al., 2003), however the emphasis of the tutor, and most of the feedback, is on problem solving. Katz et al. noticed that the human tutors tended to address issues concerning step-by-step problem solving *during* problem solving, and then used the post-problem solving reflective dialogs to elaborate on conceptual explanations and provide more abstracted solution schema. For many students using this tutor, it may be that post problem solving reflection was their first opportunity to think about more abstract conceptual issues.

In contrast, the Why2-Atlas tutor, and therefore the Itspoke tutor, was designed to directly address conceptual understanding. It tutors qualitative physics problems (as opposed to quantitative physics problems) and emphasizes conceptual explanations. The major feature of the reflective text used in this experiment, comparing different problem solving episodes, was not present in the tutoring dialogs. However the reasons for each application

of Newton's laws were discussed, and seems likely that sufficiently motivated students would be able to pick up a conceptual understanding more easily from Itspoke than from the Andes tutor. Consequently, reflection after Itspoke might be expected to have a more narrow effect on learning than did reflection after Andes.

Another difference between reflection after Itspoke and reflection in almost all other evaluated studies is shown in Table 40. In most of the other studies, students were asked to reflect explicitly, usually by answering a question or responding to some other kind of reflective prompt. As described, explicit reflection seems in general to produce more benefit than implicit reflection, probably because it ensures that the student engages the material more actively. The Itspoke study, however, required only implicit reflection, relying instead on features of the text (i.e. its cohesiveness) to encourage engagement and active abstraction.

7.2 RELATED WORK IN HUMAN TEXT PROCESSING

In this section I discuss related work in text comprehension which, like the current study, has dealt with the relationship between motivation, knowledge and cohesion.

7.2.1 Motivation, Knowledge And Learning From Text

Schiefele and Krapp (1996) investigated the effect of topic interest on recall of expository text. In their work, they define topic interest as the "relatively long term orientation of an individual towards a certain topic, or a domain of knowledge." Their topic interest questionnaire consisted of two parts, "feeling related valences" and "value related valences." The feeling related questions asked how the subject felt in relation to the topic ("bored," "interested" etc). The value related questions asked the student to rate the personal value of the topic to them on a 1 to 4 rating scale ranging from "completely true" to "not at all true" for terms like "meaningful," "unimportant" "worthless" and so forth. Note that these questions are very similar to the intrinsic value questions used in our motivation survey, as described in Section 4.4.3. The questions on Schiefele and Krapp's instrument (both feeling

and topic interest together) had a very high reliability coefficient (alpha) of .91.

They found that interest was significantly related to free recall of idea units from the text. Interest was also related to the correctness of the sequence of ideas recalled (a measure of comprehension), and to the recall of "new" idea units, which had been implied but not stated in the text (a measure of comprehension and elaboration). They also found that interest (but not prior knowledge) was significantly related to measures of active engagement with the text such as intensity of attention, elaboration and note-taking while reading. Schiefele and Krapp concluded that student interest increases the quantity of recall and also the depth of processing, and that the relationship between interest and learning was independent of prior knowledge.

The study described above was among university students. In (Schiefele, 1996) Schiefele found a similar result among 12th grade students. Students were assessed for prior knowledge of and interest in the topic of an expository text ("pre-historic people" or "television"), then asked to read the text. After reading, the students were asked to rate if certain sentences had appeared in the text. The sentences were designed to separately test three different levels of representation hypothesized in van Dijk and Kintsch's text processing theory (vanDijk and Kintsch, 1983): the verbatim representation of the superficial text structure, the propositional representation and the situation model. Results suggested that interest was negatively related to the shallow verbatim representation, but positively related to the deeper propositional representation. There were no significant correlations for the situation model representation. This result was consistent with van Dijk and Kintsch's expectation that highly interested subjects would more actively engage the text to build deeper level (propositional and situation model) representations, and that less interested readers would instead build more shallow verbatim representations.

The above study used non-physics texts, however in (Alexander et al., 1994) Alexander, Kulikowich and Schulze found significant relationships between topic knowledge, domain interest and retention from reading a physics test. Two physics texts were used in this study, with higher and lower levels of technicality. After reading, students were asked to evaluate their interest in the whole text and on each passage within the text on a 1-to-10 scale. Domain knowledge and reading comprehension were both evaluated by having students

complete fill-in-the-blank questions. The post reading comprehension questions tested recall of various facts from the text, and had very high reliability ratings. No attempt was made to evaluate different levels of representation as had been done in (Schiefele, 1996).

Alexander et al. (1994) found that both prior domain knowledge and interest were significantly related to comprehension of the physics text, however the effect of prior knowledge was larger. Domain knowledge accounted for 21% of the variance in recall, while interest explained 5% of the variance.

In a review of many published studies, Tobias (1994) concludes that although interest and prior knowledge are well correlated, some 80% of the variance in interest was not explained by prior knowledge, and could therefore affect learning independently. Based on his literature review, Tobias suggested that interest affected learning by invoking deeper types of comprehension processes and greater use of imagery, among other things.

Finally, as mentioned in Section 2, Boscolo and Mason (2003) found that both higher prior knowledge and higher interest improved text comprehension, and that these effects seemed to be additive, so that students with high interest and high prior knowledge learned significantly more from the text than other groups.

These studies, taken together, strongly suggest that both topic interest and topic knowledge affect how a text is processed, and how much of it is retained.

7.2.2 Cohesion And Learning From Text

In Section 2.4, I described work by Britton and Gulgoz (Britton and Gulgoz, 1991) , McKeown et al. (McKeown et al., 1992) and McNamara (McNamara et al., 1996) showing that, in general, increasing the cohesiveness of a text can improve how well it is understood. In addition, I reviewed work by McNamara and her colleagues (McNamara et al., 1996; McNamara and Kintsch, 1996; McNamara, 2001; O'Reilly and McNamara, 2007) showing that *low* cohesion text can actually be beneficial for certain students.

Rather than repeat that review, this section will expand on McNamara's results which suggested that textual cohesion affects the type and depth of processing achieved. As described in Section 2.4, VanDijk and Kintsch's theory of text comprehension (vanDijk and

Kintsch, 1983) holds that several different levels of representation can be generated while reading a text.

In many of McNamara's studies (e.g. (McNamara et al., 1996; McNamara and Kintsch, 1996)) low cohesion text led to improved situation model representations, as measured by improved performance on a sorting task. This was interpreted to be because the inferences necessary to process a low cohesion text had caused learning, and produced a more elaborated situation model.

However in others (e.g. (O'Reilly and McNamara, 2007; McNamara, 2001)) low cohesion only produced gains at the propositional model level, not the situation model level. This difference was attributed to the difficulty of the texts. In (O'Reilly and McNamara, 2007) the texts concerned cell mitosis, contained many unfamiliar terms, and were difficult for most students. McNamara suggests that when unable to form situation model representations from these texts, subjects fell back to a propositional level representation.

Similar difficulty in affecting the situation model occurred in the Schiefele study (Schiefele, 1996) described above. In that study interest was positively correlated with creation of a propositional representation, but not correlated with creation of a situation model representation.

7.2.3 Relation To Current Work

The results shown in the current study seem to fit the general pattern seen in related work, given certain assumptions.

In the current study, reflection was shown to help students with a middle amount of motivation, but not those with low or high motivation. It seems safe to assume that students with very low motivation are not likely to engage in active reflection from either the reflective or control condition texts, and so show no difference in learning between conditions. The results described in Section 7.2.1 show that increased motivation alone can lead to greater engagement with and learning from text. It thus seems likely that our highly motivated subjects engaged in active processing, and possibly reflection, in both the reflection and control texts, and so also show no difference in learning between conditions.

The cognitive load measures collected during the study also provide some evidence that this interpretation is correct, and that subject motivation affects engagement with the text. As described in Section 5.6.2, the reading speed measure of cognitive load was significantly different for different levels of motivation. Both middle and high motivation students had significantly higher cognitive loads than did low motivation students. This suggests that, similarly to the work reviewed above, higher motivation is also associated with more active text processing in our corpus.

The results for our middle motivation subjects more closely follow the pattern we expected based on the related work described in Sections 7.1 and 7.2.2, above. These students did learn more from reflection, and also showed a significant learning gain interaction between domain knowledge and the cohesiveness of the text. There are still two puzzling aspects to this result, however. First, low cohesion seemed to help *low* pre-testers, where in McNamara's work it helped *high* pre-testers. Second, the benefits of reflection seemed to be more in retention than in far transfer. That is, reflection showed significant learning advantages for near transfer and delayed near transfer learning, but not for far transfer or delayed far transfer learning.

I first consider the question concerning levels of domain knowledge. One possible interpretation is that the middle motivation low pre-testers in our study are like the high pre-testers in McNamara's studies. That is, they may have roughly the same levels of knowledge relative to their domains. Remember that in both this study and McNamara's work, "high" and "low" knowledge was defined relative to the distribution of pre-test scores, rather than against some absolute scale. These studies were done in different domains and with different populations of students, but it seems plausible that a below average freshman understanding of force and motion is equivalent to an above average understanding of cell mitosis (or the Vietnam war). In this case, like McNamara's high pre-testers, our low pre-testers would have a sufficiently high level of knowledge to successfully make inferences and bridge the gaps in low cohesion text.

If this interpretation is correct, then what about the high pre-testers in the current study, who would presumably be "extra-high" in one of McNamara's studies? Why were they not also helped by low cohesion text? As shown in the related work described above, high domain

knowledge has also been associated with more active processing and increased learning from text. Perhaps these students are actively engaging text no matter what its level of cohesion, but learning more from the high cohesion text, perhaps because it has a larger number of explicit propositions on which to base inference.

This analysis suggests an additional statistical test to compare cognitive load between pre-test categories. An Anova explaining the "self-report" measure of cognitive load by pre-test category (highPre or loPre) shows a significant ($p = .039$) difference between categories, with the high pre-testers having higher cognitive load than the low pre-testers over all types of text.

This additional cognitive load result supports the interpretation being built, and invites one further inference. If these students with "extra-high" domain knowledge engage the text more actively, what are they doing? Possibly they are reconciling the text with their existing domain schema, in the way suggested by Kalyuga and Ayres (2003). Thus, our results suggest a way to partially reconcile the conflicting interpretations of McNamara and Kalyuga for the cohesion reversal effect. Perhaps students with a moderate amount of domain knowledge respond to low cohesion largely by making inferences that they wouldn't make from high cohesion text. These students are represented by the low pre-testers in the current study, who may have learned more from low cohesion text, (Table 27) and who had a higher level of cognitive load (Table 29). Students with a high amount of domain knowledge, on the other hand, respond to high cohesion by actively reconciling it with their existing schema, resulting in greater learning (Table 27) and higher cognitive load (Table 29).

This analysis, while suggestive, suffers from a few shortcomings. First, interaction with learning gains shown in Table 27 is for moderately motivated subjects. The interaction with cognitive load shown in Table 29, on the other hand, is for the group of all students, and is not significant for middle motivation subjects. Second, Kalyuga's interpretation suggests that schema reconciliation is extraneous load, and should produce less learning than the low cohesion text. Our high pre-testers instead learned more from high cohesion text. I argue in Section 8.3 that this suggests productive avenues for further research.

Next I consider the question about near- and far-transfer learning. The work in abstraction reviewed in Section 2.2 led us to believe that successfully processing an abstrac-

tive/reflective text would lead to useful mental abstractions, and so to better transfer of learning. So, given that the reflective text used in this study significantly improved learning, why was the learning only "near" transfer and "delayed near" transfer?

There are (at least) two possibilities. The first is that, similarly to McNamara's students who read about cell mitosis, our middle motivation students did not successfully form a situation model representation of the text, and so fell back on their propositional level representations. This explanation is consistent with McNamara's results for difficult texts, but does not sit easily with the suggestion just made, that these students had a relatively high level of domain knowledge.

Another possibility is that these students *did* form a situation model representation of the text, but that it was inadequate for solving far-transfer questions. This explanation seems more satisfying for several reasons. First, the *delayed* post-test results suggest that the representation being used was the deeper situation model, and not one of the more shallow representations which are thought to decay more rapidly. Second, the cognitive load results presented in Section 5.6.2 suggest that these middle motivation students did process the text more actively, relative to the more poorly motivated students. Their higher level of cognitive load makes it easier to believe that they were successfully forming a situation model from it. If this is true, then it suggests that there are significant differences between even a deep textual representation and whatever operationalized schema is necessary to solve transfer problems.

7.3 RELATED WORK IN COMPUTATIONAL LINGUISTICS

In this section I first briefly review work from the computational linguistics community which has explored a number of ways to measure cohesion [1] in text. Following that, I describe additional work that has measured cohesion in dialog. Finally, I compare both

[1] Because these metrics all measure properties of text, rather than of a reader's understanding, I use the term "cohesion" throughout. Some of the original papers use a different convention. For example, some of them use "cohesion" for local syntactic relations in the text, and "coherence" for longer range rhetorical or semantic relations in the text.

these communities to the dialog cohesion measurements described in Section 6.

7.3.1 Measures Of Cohesion In Text

The various approaches to measuring cohesion in text can be divided into approaches based on centering theory, approaches using corpus based similarity measures, and approaches using semantic resources such as WordNet. I will give a few examples of each in turn.

Barzilay and Lapata (2005) measure textual cohesion by modeling the focus state of a hypothetical reader. This approach is inspired by centering theory (Grosz et al., 1995), which was intended to model a reader's attentional state, as described in Grosz and Sidner's theory of discourse structure (Grosz and Sidner, 1986). This approach creates a grid showing the transitions certain discourse entities make between syntactic functions in successive sentences. It models the "smoothness" of the writing as a function of the syntactic roles each entity assumes. The scores were used to rank texts by coherence.

Miltsakaki and Kukich (Miltsakaki and Kukich, 2004) describe an improvement in an essay scoring system which also makes use of a measure of cohesion based on centering theory. They automatically identify one type of transition, the "rough shift," and find it is helpful in identifying poorly structured essays.

Work in text segmentation (Hearst, 1994) has used cohesion to identify topic boundaries in text. Cohesion was measured by counting the number of repeating words between adjacent sentences, and topic boundaries were identified in places where sentence-to-sentence similarity dropped.

Corpus based approaches to cohesion measure similarity between words by looking at their distributions in a corpus. For example, latent semantic analysis (LSA) and its variations can be thought of as doing a type of implicit inference (Foltz et al., 1998). It infers that two words are similar if they tend to be used in similar contexts. Foltz et al. (1998) measure the cohesiveness of a text by using LSA to measure the distributional similarity of words in consecutive sentences. They found that this measure of document cohesiveness was significantly correlated with human judgments.

Higgins et al. (2004) describe another method of measuring the cohesion of an essay.

They identify various discourse segments commonly present in a good essay, such as thesis, support and conclusion. They then use vector based methods to measure the semantic relatedness of each sentence to its segment, as well as to a few other things such as the original question.

Lapata and Barzilay (2005) combine several of the approaches described above to measure textual cohesion. They use both the entity-grid syntactic measure of coherence inspired by centering theory, and also several semantic measures such as the LSA measure used by Foltz. They find that the two approaches are complimentary, and the combination models perform best on a cohesion scoring task.

Semantic Resource Approaches to measuring textual cohesion use resources such as WordNet or WikiPedia. They determine the similarity between two words by examining the path between them in the resource. These methods commonly produce a "similarity" (or inversely, a "distance") number like the LSA methods, but can also produce link type information. For example, they can tell if the words connected by hyponymy or synonymy relations, how many direction switches occur in the path, and so on. This additional path information allows us to place them more toward coherence than the pure corpus based measures.

Resnik (1995) finds the similarity between words as the information content of their nearest common subsuming concept in an "is-a" (hyponymy) hierarchy. The information content of the subsuming concept is learned from a corpus. Using additional information from a corpus addresses the problem of link density in semantic resources. Simply counting links does not give a good way to compare semantic distances, because the different portions of the semantic hierarchy might be unevenly developed.

7.3.2 Measures Of Cohesion In Dialog

Section 7.3.1 above, describes how latent semantic analysis (LSA) has been used to measure sentence to sentence cohesion in text. Olney and Cai (2005) extend this measure for use in dialog. They use a variation of LSA to segment a corpus of tutorial dialogs into cohesive topics. They use orthonormal projection to measure the extent to which each successive

utterance contains new or old content compared to the dialog so far. This measure of utterance-to-utterance similarity is used to predict the topic boundaries between different tutored problems in the corpus.

Other work examining the cohesiveness of tutorial dialog has been done by the AutoTutor group at the University of Memphis. In (Graesser et al., 2007), they use the CohMetrix (Graesser et al., 2004) cohesion analysis tool to analyze the cohesiveness of tutor and student dialog contributions along many dimensions. Using this tool, they show differences in cohesiveness between tutorial dialog and other types of discourse.

Work examining tutorial dialog cohesion has been done by Jeon and Azevedo (2007, 2008) in the domain of learning about the human circulatory system with hypermedia. In (Jeon and Azevedo, 2008), they used the CohMetrix tool to measure cohesion in a corpus of tutor-student dialogs, from which the tutor utterances had been removed. Among other things, they found higher latent semantic analysis (LSA) scores for the utterances of students who had improved their mental model of the circulatory system, compared to those who had not. In (Jeon and Azevedo, 2007), they find significant differences in cohesion between transcripts of students engaged in self regulated learning, vs. in externally regulated tutoring.

7.3.3 Relation To Current Work

In Section 6, I describe a method of evaluating learning during tutorial dialog by measuring the cohesion between adjacent tutor and student utterances. This method is similar to several of the textual measures described in Section 7.3.1. My lexical measure counts the repetition of words between tutor and student, and so is very similar to the text-tiling method (Hearst, 1994) used to segment expository text. My semantic measure of cohesion counts the repetition of words that are different but have similar meanings, where semantic similarity is measured using path distance in the WordNet hierarchy. This was inspired by the corpus similarity measures described above, for example (Resnik, 1995). However, where these measures were designed for text, the measures described in Section 6 were adapted for tutorial dialog.

Work described in Section 7.3.2 does measure cohesion in tutorial dialog, and this work

has interesting differences from my cohesion measurement described in Section 6. The work using CohMetrix described above generally treats the tutorial dialog as a text. It separates out all the student utterances, or all the tutor utterances, and uses CohMetrix to measure the cohesiveness within that group. My work, in contrast, measures cohesion between adjacent tutor and student utterances.

CohMetrix, however, provides a much greater variety of cohesion measures than I have developed, although there are similarities. For example, our semantic measures are similar in spirit, but where they use LSA to gauge the distributional similarity between two turns, I use a WordNet similarity metric to locate specific pairs of similar words between turns. Their "argument overlap" metric is also very similar to my lexical reiteration measure.

The work of Olney and Cai (Olney and Cai, 2005), described above, may be the closest to my tutorial dialog methods. They also measure similarity between successive utterances in tutorial dialog. However, they use their measure to predict topic boundaries, whereas my measure was used to predict learning from tutoring.

8.0 CONTRIBUTIONS AND FUTURE WORK

In this section I first review contributions made by the current work, then discuss a few of its limitations. Based on the contributions and limitations, I then discuss a few avenues for future work.

8.1 CONTRIBUTIONS

This thesis work has made contributions to each of the three related communities whose work was described in Section 7.

This work contributes to research in reflection in Intelligent Tutoring Systems, which was outlined in Section 7.1, in several ways. First, by confirming Hypothesis One (Section 5.3) this work has shown that reading an abstractive-reflective text after tutoring can improve both immediate and delayed measures of learning after tutoring in *qualitative* physics, for moderately motivated students. This result extends other work described in Section 7.1.2 which showed that reflection can benefit qualitative learning after *quantitative* tutoring.

Second, it shows that reflective prompts such as those implemented for this study, which compare previous problem solving episodes, are effective even when they require only *implicit* reflection on the part of the student. In addition, this work's use of implicit reflection has contributed results for a region of study design space that was previously unexplored (see Table 40).

Third, by showing success while focusing on abstractive comparisons of previous problem solving episodes, this work adds to evidence that abstraction is a major reason for reflection's impact on learning.

Forth, the sensitivity of this intervention to motivation level suggests that motivation is a significant factor in learning from ITS's. It implies that student models should include estimates of motivation, and that these estimates could be used to determine for which students reflective prompts would be necessary or useful.

This work also makes contributions to the field of human text processing research, which was outlined in Section 7.2. By confirming Hypotheses Two (Section 5.4) and Three (Section 5.5) for middle motivation subjects, this work has shown that the cohesiveness of a reflective text can affect how much students learn from reading it, and that this effect interacts with student knowledge. Furthermore, results gained from testing the three primary hypotheses contribute evidence that the benefit of reflection, and the effects of textual cohesion, are both strongest for students with a "middle" amount of motivation.

In addition, the cognitive load results reported in Section 5.6.2 suggest *why* this intervention was most effective for middle motivation students. Higher cognitive load for high and middle motivation readers, relative to low motivation readers, suggest that motivation, at high levels, causes more active processing of the text.

These results add to evidence of significant interactions among domain knowledge, motivation and textual cohesion that has been reported by other researchers. This work also extends those findings into a new domain: reflective tutorial text.

Finally, the work described in Section 6 should be of interest to researchers in the field of dialog cohesion, which was outlined in Section 7.3.2. It has demonstrated a method for extending textual measures of cohesion for use in dialog, and shown that they correlate with what one dialog participant, the student, is learning. The initial study in the series (Ward and Litman, 2006) was the first work to use measures of tutorial dialog cohesion to predict learning, a result which has now been replicated in several corpora of tutoring dialogs.

8.2 LIMITATIONS OF THE STUDY

In addition to the contributions described above, the work described in this thesis suffers from several limitations, a few of which I describe here.

A major limitation of this study is the low reliability of the cognitive load measures used, as described in Section 5.1.3. This weakens both results for secondary Hypotheses Four and Five, and also judgments about the underlying mechanisms behind the interactions that were discovered.

Another limitation concerns interpreting the effect of student motivation. Results in this study can't conclusively determine if the effect of student motivation is because of differences in model encoding during reading, or because of more active elaboration of the model during problem solving. However since shallower representations are generally thought to decay more rapidly than the deeper ones, the fact that some of the results in this study were for delayed retention suggests that more than verbatim representations may have been involved.

Another limitation of the study is that it cannot tease apart the effect of text comprehension on transfer. That is, although the delayed post-test results provide a hint (as described in Section 7.2.3), we cannot tell for certain if far-transfer learning was hindered because students failed to form an adequate situation model of the text, or because even a well formed situation model is not sufficient for problem solving.

Finally, this study did not allocate subjects among conditions with reference to their motivation level. Post-hoc categorization of subjects by their motivation scores, therefore, led to an uneven distribution of subjects among cells, and some low counts, as shown in Tables 14, 15 and 16.

8.3 FUTURE WORK

This work has left a number of issues unsettled, and suggested other interesting areas for exploration. In this section I will first mention two areas of future work suggested by the current study's shortcomings, and one future project suggested by the current study's success.

This thesis work has shown significant interactions between cohesion and domain knowledge, with effects on both learning gains and cognitive load. This suggests that domain knowledge could be used to personalize the cohesiveness of expository texts. This study, as well as previous research, has defined high and low knowledge using a mean or median

split on the current subject population. However this approach makes it difficult to know beforehand, based just on the current pre-test distribution, if a particular student has "high" enough knowledge to need a low cohesion text. For example, in the current study, students who were "low" knowledge relative to the distribution of pre-test scores may have actually been "high" enough, relative to the domain, to benefit from low cohesion text. A very useful avenue for future research, therefore, would be to define "high" and "low" pre-testers with reference to concepts gleaned from a cognitive task analysis of the tutored problems and test questions. Such an analysis would allow us to more closely determine the relationship between the presence of concepts in prior knowledge and the effect of making them explicit or implicit in text.

Another issue raised by these results is the exact relationship between the mental structures built when comprehending text, and the abstract schema thought to be necessary for far transfer problem solving. Cognitive load results suggest that our middle motivated students were processing the text more deeply than the least motivated students, while their performance on the delayed post-test suggests that they were building the type of deep level model which is thought to decay more slowly. However, this model did not seem to be adequate to improve performance on far transfer questions which were dissimilar to the tutored problems. This suggests that interesting work could be done which measured both the students' text representations and post-test performance, and attempted to find relationships between the two. This would help clear up whether the difficulty lay in comprehending the text at a deep enough level, or in translating that comprehension into useful problem solving knowledge.

This work also raises issues about the relationship between text and dialog in tutoring, and how they could be best combined in an Intelligent Tutoring System. An interactive and responsive tutoring system is often thought to be the holy grail of tutoring research, largely because interactivity is expected to cause more knowledge creation on the part of the student. However, not all the knowledge covered during interactive tutoring can be generated by the student. At each point in the dialog, the tutor must make the decision whether to "elicit" that piece of knowledge, or to simply "tell" it. It has been shown that sophisticated policies governing this "elicit-vs-tell" decision can improve learning (Chi et al.,

2010). Chi et al. (2010) show that the optimal learned ratio of "elicit" vs "tell" actions can be different for different knowledge components, and that it is often effective for a tutor to "tell" rather than continuously "eliciting."

In fact, other evidence suggests that it is sometimes best for the tutor to use a long series of "tells" without eliciting any knowledge at all from the student. As observed by D'Mello and colleagues (D'Mello et al., 2010), even in interactive tutoring it is sometimes necessary for the tutor to enter "lecture" mode. In cases where the student lacks the knowledge needed to respond effectively to scaffolding, the tutor is forced to simply deliver the missing information. This raises the question of how that missing information should be delivered, as a mini-lecture, embedded in the dialog, or as a text to be read?

The inclination to expect that interactive dialog will be better even in this situation may not be well founded. Most of the benefits of interactivity will clearly be less present in "lecture" mode. In addition, there are several reasons to expect that text may actually be better. For example, students may read at their own pace, without feeling rushed to fulfill a dialog obligation to respond periodically. Text can also be re-read, which is difficult in a dialog situation. In addition, because of its ubiquity in schools, text is well studied. Researchers have found many features of text (such as its cohesion, its title structure, or the presence and placement of its graphics) which can make it more effective for certain students.

These considerations suggest that finding the best way to combine text and interactive dialog in a tutoring system is an important topic for research, although it has so far received little attention. It should also be noted that text and interactive tutoring are not mutually exclusive, but could be combined in many ways. For example, the tutor could switch to text mode when it decides the next series of points should be delivered primarily in "tells", as suggested above. Or, it could present a low cohesion text, then use interactive dialog to force the student to address the gaps in the text.

Another option for combining text and tutoring is suggested by the use of a reflective text in the current study. Under this option, a tutor might alternate between presenting some form of tutoring and presenting a text. A major purpose of the tutoring sessions would be interactive testing, which would be used to build a model of student knowledge and motivation. Among other measures, dialog measures of cohesion, as described in Section 6

could be used during the tutoring phase to measure learning. The post-tutoring text would then be generated dynamically to have the specific content and cohesive structure deemed best for that student. The student would read the text then return to tutoring for further evaluation.

Text generation is a sub-field of Natural Language Processing, which studies how to generate text output from computer-internal semantic representations. The output is typically generated in several stages, for example first a "strategic" stage which determines the general content and structure of the message, and then a "tactical" stage which translates the message into natural language output (McKeown, 1985). These choices are often made with reference to various theoretical models of discourse. For example, rhetorical structure theory (Mann and Thompson, 1988) is often used to select a message structure that advances the appropriate discourse goal, while a model of the reader's focus state (Grosz et al., 1995) is sometimes used to avoid jarring "rough shifts" in the surface form.

This type of text generation mechanism has been used to generate tutor utterances in tutorial dialog (Freedman, 1996), and also to generate web pages in dynamic hypertext systems (O'donnell et al., 2001). However, it does not seem to have been used to generate individualized instructional text during tutoring, in the way imagined above.

9.0 ACKNOWLEDGEMENTS

I am deeply grateful to my advisor, Diane Litman, and to my committee for helpful advice at every point in this research. I have also gotten invaluable help from many others along the way, including (but not limited to) Tessa Warren, Bob Hausmann, Chas Murray, Mihai Rotaru, Joel Tetreault and the Itspoke group. Their collective help and advice is responsible for much that is good about this work. I generated all the flaws by myself, probably over their objections.

Finally, I am deeply grateful to my wife Lauren, and daughter Anne. Without their support my passage through graduate school would have been unthinkable.

This work was supported by the NSF (Grant # 0631930 and 0914615) as well as by Graduate Student Researcher funding from both the Learning Research and Development Center (LRDC) and the Intelligent Systems Program (ISP) at the University of Pittsburgh.

The previous research in dialog cohesion which is reported in Section 6 was supported by the NSF (0325054) and ONR (N00014-04-1-0108).

APPENDIX A

EXPERIMENTAL TEXTS

This section reproduces the experimental texts used for both the pre- and post-tutoring readings. These readings were all presented block-by-block in the Linger interface, and the block dividers and block ids are indicated in the texts. For example "Reading Block Divider: baseline1 e" indicates that the following text starts a new block, and that the block id is "baseline1." The letter following the block id is a code controlling what test Linger presents following the block.

A.1 BASELINE AND INTRODUCTORY PRE-READING TEXTS

Reading Block Divider: lingerIntro a

In a minute you will read some introductory material about physics. First, however, is a short text to familiarize you with the reading environment we will be using. This is a "tap-to-read" environment. Please read at a comfortable speed. When you get to the end of a block of text, tap 'return' for the next block. Occasionally, after reading a block of text, a "reading difficulty" question will appear. This question will ask you to rate, on a 1-to-7 scale, how difficult the previous passage had been to read. 1 means "very easy" and 7 means "very hard." Please complete the following practice reading now.

Reading Block Divider: baseline1 e

When Farmer Oak smiled, the corners of his mouth spread till they were within an

unimportant distance of his ears, his eyes were reduced to chinks, and diverging wrinkles appeared round them, extending upon his countenance like the rays in a rudimentary sketch of the rising sun.

His Christian name was Gabriel, and on working days he was a young man of sound judgment, easy motions, proper dress, and general good character. On Sundays he was a man of misty views, rather given to postponing, and hampered by his best clothes and umbrella: upon the whole, one who felt himself to occupy morally that vast middle space of Laodicean neutrality which lay between the Communion people of the parish and the drunken section, – that is, he went to church, but yawned privately by the time the congregation reached the Nicene creed,- and thought of what there would be for dinner when he meant to be listening to the sermon. Or, to state his character as it stood in the scale of public opinion, when his friends and critics were in tantrums, he was considered rather a bad man; when they were pleased, he was rather a good man; when they were neither, he was a man whose moral colour was a kind of pepper-and-salt mixture.

Reading Block Divider: baseline2 e

Since he lived six times as many working-days as Sundays, Oak's appearance in his old clothes was most peculiarly his own – the mental picture formed by his neighbours in imagining him being always dressed in that way. He wore a low-crowned felt hat, spread out at the base by tight jamming upon the head for security in high winds, and a coat like Dr. Johnson's; his lower extremities being encased in ordinary leather leggings and boots emphatically large, affording to each foot a roomy apartment so constructed that any wearer might stand in a river all day long and know nothing of damp – their maker being a conscientious man who endeavoured to compensate for any weakness in his cut by unstinted dimension and solidity.

Reading Block Divider: baseline3 e

Mr. Oak carried about him, by way of watch,- what may be called a small silver clock; in other words, it was a watch as to shape and intention, and a small clock as to size. This instrument being several years older than Oak's grandfather, had the peculiarity of going either too fast or not at all. The smaller of its hands, too, occasionally slipped round on the pivot, and thus, though the minutes were told with precision, nobody could be quite certain

of the hour they belonged to. The stopping peculiarity of his watch Oak remedied by thumps and shakes, and he escaped any evil consequences from the other two defects by constant comparisons with and observations of the sun and stars, and by pressing his face close to the glass of his neighbours' windows, till he could discern the hour marked by the green-faced timekeepers within. It may be mentioned that Oak's fob being difficult of access, by reason of its somewhat high situation in the waistband of his trousers (which also lay at a remote height under his waistcoat), the watch was as a necessity pulled out by throwing the body to one side, compressing the mouth and face to a mere mass of ruddy flesh on account of the exertion, and drawing up the watch by its chain, like a bucket from a well.

Reading Block Divider: physicsIntro a

Thank you for completing the warm-up reading. The next reading is meant to introduce you to the basic physics concepts that you will use later while working through the set of physics problems.

Reading Block Divider: veloc e

Velocity:

In everyday language, we can use the words speed and velocity interchangeably. In physics, we make a distinction between the two. Very simply, the difference is that velocity is speed in a given direction. We say a car travels at 1.1 m/s (meters per second), we are specifying its speed. But if we say a car moves at 1.1 m/s to the north, we are specifying its velocity.

Scalar vs Vector quantities:

Quantities that require both magnitude and direction for a complete description are called vector quantities. Quantities that can be specified using only magnitude are called scalar quantities. Displacement and velocity are vector quantities, while speed and volume are scalar quantities.

Text Divider: constVel e Constant velocity:

From the definition of velocity it follows that to have a constant velocity requires both constant speed and constant direction. Constant direction means that the motion is in a straight line; the object‘s path does not curve at all.

Text Divider: accel e Acceleration:

We can change the velocity of an object either by changing its speed, by changing its direction of motion, or by changing both. The rate at which velocity is changing (either increasing or decreasing) is called the acceleration. Because acceleration is a rate, it is a measure of how fast the velocity is changing per unit of time. Acceleration = change in velocity / time interval.

Text Divider: forces e Forces:

A force is any push or pull. Two objects are always involved whenever force is exerted. When I push a car, I am one object and the car is the other. Physicists say that the force 'acts on' one of the objects and is 'due to' the other. Thus, when I push a car, the force acts on the car and is due to me.

Contact forces:

Forces are put into two categories: contact forces and field forces. If the two objects are pressing against each other, they are involved in a contact interaction, then the force is called a contact force. Contact forces only exist when the two objects are in physical contact.

Field forces:

Non-contact forces are called field forces. When a planet pulls on an object near it, the pull is called a gravitational force. A gravitational force exists even when the planet and the object are not touching.

Text Divider: newtOne e Newton‘s First Law:

Newton's First Law is: 'Every body continues in its state of rest, or of motion in a straight line at constant speed, unless it is compelled to change that state by forces exerted upon it.' An object at rest will remain at rest until a force is applied to it.

Now consider an object in motion. If you slide a hockey puck along the surface of a city street, the puck quite soon comes to rest. If you slide it along ice, it slides for a longer distance. This is because the friction force on it on the surface of ice is very small. If friction is absent, it slides with no loss in speed. We thus conclude that in the absence of applied forces, a moving object will move in a straight line with constant speed indefinitely.

Text Divider: mass e Mass:

Kick an empty tin can and it moves. Kick a tin can filled with solid lead, and you‘ll only hurt your foot. Even though it has the same volume, the lead-filled can has greater

resistance to motion than the empty can because it has greater mass. The mass of an object is a measure of its resistance to changing its motion.

Text Divider: massNotWeight e Mass is not weight:

Now that you have an understanding of field forces, you are ready to understand the distinction between mass and weight. Mass is an intrinsic property of an object; it is determined by the actual material in the body. It depends only on the number and kind of atoms that compose it. Weight is a measure of the gravitational force that acts on the body, and hence depends on where the object is located.

The amount of material in a particular stone is the same whether the stone is located on the earth, on the moon or in outer space. Hence, its mass is the same in any of these locations.

But the weight of the stone would be very different on the earth, on the moon, and in outer space. On the surface of the moon, the stone would have only one-sixth of the weight it has on the earth. This is because the acceleration due to gravity is only one-sixth as strong on the moon as compared to that on the earth. If the stone were in a gravity-free region of space, its weight would be zero. Its mass, on the other hand, would remain the same everywhere.

Near a planet, the mass of an object is directly proportional to the magnitude of its weight (and weight is gravitational force). The constant of proportionality is called g, so W = mg, where W is the magnitude of the object‘s weight, m is the object‘s mass and g is a constant that depends on the planet. For Earth, g = 9.8. For the moon, g = 1.6.

Text Divider: forceAcc e Forces produce acceleration:

The key idea behind Newton‘s first law is that it takes a force acting on an object in order to cause a change its motion. Consider an object at rest, such as a hockey puck on a smooth, nearly frictionless ice. Push it with a stick and it begins to move. Its velocity changed from zero in the beginning to some value at the end of the push. When the stick is no longer in contact with the puck, the puck moves at constant velocity. Apply another force by striking it with the stick again, and the motion changes. Again the puck has accelerated. Force produces acceleration.

Text Divider: netForce e Net force:

Often, there is more than one force acting on an object. The combination of all the forces that act on an object is called the 'net force.' It is the net force that produces the acceleration of an object. For instance, suppose you attach two threads to a puck. If you pull on one thread, the puck accelerates. If your friend also pulls on the second thread in the same direction, the net force increases and the puck accelerates more.

On the other hand, if your friend pulls with the same force as you, but in the opposite direction, then the two pulls cancel each other, the net force is zero, and the puck does not accelerate. Thus, acceleration is due to the net force on an object, which is the sum of all the individual forces acting on the object.

Text Divider: newtTwo e Newton‘s second law:

Push on an empty shopping cart, then push equally hard on a heavily loaded shopping cart, and you‘ll produce much less acceleration in the second case. This is because acceleration depends on the mass of the object being pushed. For objects of greater mass, we find smaller accelerations for the same force. Newton's second law formalizes it this way: 'The acceleration of a body is directly proportional to the magnitude of the net force acting on it and inversely proportional to its mass, and the acceleration is in the same direction as the net force.' That is, acceleration = net force / mass.

From this relationship, we can see that if the net force that acts on an object is doubled, the acceleration will be doubled. Suppose instead that the mass is doubled. Then the acceleration will be halved. If both the net force and the mass are doubled, then the acceleration will be unchanged.

Text Divider: normForce e Normal force:

How many forces act on your book as it lies motionless on the table? Don‘t say one, its weight. If that were the only force acting on it, you‘d find it accelerating. The fact that it is at rest, and not accelerating, is evidence that net force on it is zero. So, another force must be acting in opposite direction. The other force is called 'the normal force due to the table acting on the book.' The table actually pushes up on the book with the same amount of force that the book presses down. If the book is to be at rest, the sum of the forces acting on it must balance to zero.

Text Divider: falling e Falling through air:

Drop a stone and it falls. Does it accelerate while falling? We know it starts from a rest position and gains speed as it falls. We know this because it would be safe to catch if it fell a meter or two, but not from the top of a tall building. Thus, the stone must gain more speed during the time it drops from a building than during the shorter time it takes to drop one meter. This gain in speed indicates that the stone accelerates as it falls.

Gravitational force causes the stone to fall downward once it is dropped. Air resistance tends to slow it down, but if the stone is still moving slowly, then air resistance can be ignored, and the gravitational force is the only force acting on the stone. Whenever gravitational force is the only force acting on an object, the object is said to be in free fall.

Text Divider: newtThree e Newton‘s third law:

In the simplest sense, a force is a push or a pull. Looking closer, however, we find that a force is not a thing in itself, but is due to the interaction between one thing and another. One force is called the action force. The other is called the reaction force. It doesn‘t matter which force we call action and which we call reaction. The important thing is that neither force exists without the other. The action and reaction forces make up a pair of forces.

In every interaction, forces always occur in pairs. For example, in walking across the floor you push against the floor, and the floor in turn pushes against you. Likewise, the tires of a car push against the road, and the road in turn pushes back on the tires. In swimming you push the water backward, and at the same time the water pushes you forward. There is a pair of forces acting in each instance.

Text Divider: axes e Coordinate axes:

When we deal with vectors it is necessary to define your coordinate axes so as to define the directions of the vector quantities. Coordinate axes are two mutually perpendicular axes, which we will refer to as the x- and the y-axes. Often we are trying to analyze the motion of an object that moves both horizontally and vertically, such as a cannon ball shot at an angle. If we choose the coordinate axes so that the x-axis is horizontal and the y-axis is vertical, the analysis is much easier.

Text Divider: compVec e Components of vectors:

Any single vector can be regarded as the sum of two vectors, each of which acts on the body in some direction other than that of the given vector. These two vectors are known as

the components of the given vector that they replace.

A man pushing a lawnmower applies a force that pushes the machine forward and also against the ground. In Figure 6-10, vector F represents the force applied by the man. We can separate this force into two components. Vector Y is the vertical component, which is the downward push against the ground. Vector X is the horizontal component, which is the forward force that moves the lawnmower.

The rule for finding the vertical and horizontal components of any vector is relatively simple, and is illustrated in Figure 6-11. A vector V is drawn in the proper direction to represent the force, velocity or whatever vector is in question (Figure 6-11 left). Then vertical and horizontal lines are drawn at the tail of the vector (Figure 6-11 right). A rectangle is drawn that encloses the vector V in such a way that V is the diagonal and the sides of the rectangle are the desired components. We see that the components of the vector V are then represented in the direction and magnitude of the vectors X and Y.

A.2 "READ-AGAIN" CONTROL TEXT

Velocity:

In everyday language, we can use the words speed and velocity interchangeably. In physics, we make a distinction between the two. Very simply, the difference is that velocity is speed in a given direction. We say a car travels at 1.1 m/s (meters per second), we are specifying its speed. But if we say a car moves at 1.1 m/s to the north, we are specifying its velocity.

Scalar vs Vector quantities:

Quantities that require both magnitude and direction for a complete description are called vector quantities. Quantities that can be specified using only magnitude are called scalar quantities. Displacement and velocity are vector quantities, while speed and volume are scalar quantities.

Constant velocity:

From the definition of velocity it follows that to have a constant velocity requires both

constant speed and constant direction. Constant direction means that the motion is in a straight line; the object‘s path does not curve at all.

Acceleration:

We can change the velocity of an object either by changing its speed, by changing its direction of motion, or by changing both. The rate at which velocity is changing (either increasing or decreasing) is called the acceleration. Because acceleration is a rate, it is a measure of how fast the velocity is changing per unit of time. Acceleration = change in velocity / time interval.

Forces:

A force is any push or pull. Two objects are always involved whenever force is exerted. When I push a car, I am one object and the car is the other. Physicists say that the force 'acts on' one of the objects and is 'due to' the other. Thus, when I push a car, the force acts on the car and is due to me.

Contact forces:

Forces are put into two categories: contact forces and field forces. If the two objects are pressing against each other, they are involved in a contact interaction, then the force is called a contact force. Contact forces only exist when the two objects are in physical contact.

Field forces:

Non-contact forces are called field forces. When a planet pulls on an object near it, the pull is called a gravitational force. A gravitational force exists even when the planet and the object are not touching.

Newton‘s First Law:

Newton's First Law is: 'Every body continues in its state of rest, or of motion in a straight line at constant speed, unless it is compelled to change that state by forces exerted upon it.' An object at rest will remain at rest until a force is applied to it.

Now consider an object in motion. If you slide a hockey puck along the surface of a city street, the puck quite soon comes to rest. If you slide it along ice, it slides for a longer distance. This is because the friction force on it on the surface of ice is very small. If friction is absent, it slides with no loss in speed. We thus conclude that in the absence of applied forces, a moving object will move in a straight line with constant speed indefinitely.

Mass:

Kick an empty tin can and it moves. Kick a tin can filled with solid lead, and you'll only hurt your foot. Even though it has the same volume, the lead-filled can has greater resistance to motion than the empty can because it has greater mass. The mass of an object is a measure of its resistance to changing its motion.

Mass is not weight:

Now that you have an understanding of field forces, you are ready to understand the distinction between mass and weight. Mass is an intrinsic property of an object; it is determined by the actual material in the body. It depends only on the number and kind of atoms that compose it. Weight is a measure of the gravitational force that acts on the body, and hence depends on where the object is located.

The amount of material in a particular stone is the same whether the stone is located on the earth, on the moon or in outer space. Hence, its mass is the same in any of these locations.

But the weight of the stone would be very different on the earth, on the moon, and in outer space. On the surface of the moon, the stone would have only one-sixth of the weight it has on the earth. This is because the acceleration due to gravity is only one-sixth as strong on the moon as compared to that on the earth. If the stone were in a gravity-free region of space, its weight would be zero. Its mass, on the other hand, would remain the same everywhere.

Near a planet, the mass of an object is directly proportional to the magnitude of its weight (and weight is gravitational force). The constant of proportionality is called g, so W = mg, where W is the magnitude of the object's weight, m is the object's mass and g is a constant that depends on the planet. For Earth, g = 9.8. For the moon, g = 1.6.

Forces produce acceleration:

The key idea behind Newton's first law is that it takes a force acting on an object in order to cause a change its motion. Consider an object at rest, such as a hockey puck on a smooth, nearly frictionless ice. Push it with a stick and it begins to move. Its velocity changed from zero in the beginning to some value at the end of the push. When the stick is no longer in contact with the puck, the puck moves at constant velocity. Apply another force

by striking it with the stick again, and the motion changes. Again the puck has accelerated. Force produces acceleration.

Net force:

Often, there is more than one force acting on an object. The combination of all the forces that act on an object is called the 'net force.' It is the net force that produces the acceleration of an object. For instance, suppose you attach two threads to a puck. If you pull on one thread, the puck accelerates. If your friend also pulls on the second thread in the same direction, the net force increases and the puck accelerates more.

On the other hand, if your friend pulls with the same force as you, but in the opposite direction, then the two pulls cancel each other, the net force is zero, and the puck does not accelerate. Thus, acceleration is due to the net force on an object, which is the sum of all the individual forces acting on the object.

Newton‘s second law:

Push on an empty shopping cart, then push equally hard on a heavily loaded shopping cart, and you‘ll produce much less acceleration in the second case. This is because acceleration depends on the mass of the object being pushed. For objects of greater mass, we find smaller accelerations for the same force. Newton's second law formalizes it this way: 'The acceleration of a body is directly proportional to the magnitude of the net force acting on it and inversely proportional to its mass, and the acceleration is in the same direction as the net force.' That is, acceleration = net force / mass.

From this relationship, we can see that if the net force that acts on an object is doubled, the acceleration will be doubled. Suppose instead that the mass is doubled. Then the acceleration will be halved. If both the net force and the mass are doubled, then the acceleration will be unchanged.

Normal force:

How many forces act on your book as it lies motionless on the table? Don‘t say one, its weight. If that were the only force acting on it, you‘d find it accelerating. The fact that it is at rest, and not accelerating, is evidence that net force on it is zero. So, another force must be acting in opposite direction. The other force is called 'the normal force due to the table acting on the book.' The table actually pushes up on the book with the same amount of

force that the book presses down. If the book is to be at rest, the sum of the forces acting on it must balance to zero.

Falling through air:

Drop a stone and it falls. Does it accelerate while falling? We know it starts from a rest position and gains speed as it falls. We know this because it would be safe to catch if it fell a meter or two, but not from the top of a tall building. Thus, the stone must gain more speed during the time it drops from a building than during the shorter time it takes to drop one meter. This gain in speed indicates that the stone accelerates as it falls.

Gravitational force causes the stone to fall downward once it is dropped. Air resistance tends to slow it down, but if the stone is still moving slowly, then air resistance can be ignored, and the gravitational force is the only force acting on the stone. Whenever gravitational force is the only force acting on an object, the object is said to be in free fall.

Newton‘s third law:

In the simplest sense, a force is a push or a pull. Looking closer, however, we find that a force is not a thing in itself, but is due to the interaction between one thing and another. One force is called the action force. The other is called the reaction force. It doesn‘t matter which force we call action and which we call reaction. The important thing is that neither force exists without the other. The action and reaction forces make up a pair of forces.

In every interaction, forces always occur in pairs. For example, in walking across the floor you push against the floor, and the floor in turn pushes against you. Likewise, the tires of a car push against the road, and the road in turn pushes back on the tires. In swimming you push the water backward, and at the same time the water pushes you forward. There is a pair of forces acting in each instance.

Coordinate axes:

When we deal with vectors it is necessary to define your coordinate axes so as to define the directions of the vector quantities. Coordinate axes are two mutually perpendicular axes, which we will refer to as the x- and the y-axes. Often we are trying to analyze the motion of an object that moves both horizontally and vertically, such as a cannon ball shot at an angle. If we choose the coordinate axes so that the x-axis is horizontal and the y-axis is vertical, the analysis is much easier.

Components of vectors:

Any single vector can be regarded as the sum of two vectors, each of which acts on the body in some direction other than that of the given vector. These two vectors are known as the components of the given vector that they replace.

A man pushing a lawnmower applies a force that pushes the machine forward and also against the ground. In Figure 6-10, vector F represents the force applied by the man. We can separate this force into two components. Vector Y is the vertical component, which is the downward push against the ground. Vector X is the horizontal component, which is the forward force that moves the lawnmower.

The rule for finding the vertical and horizontal components of any vector is relatively simple, and is illustrated in Figure 6-11. A vector V is drawn in the proper direction to represent the force, velocity or whatever vector is in question (Figure 6-11 left). Then vertical and horizontal lines are drawn at the tail of the vector (Figure 6-11 right). A rectangle is drawn that encloses the vector V in such a way that V is the diagonal and the sides of the rectangle are the desired components. We see that the components of the vector V are then represented in the direction and magnitude of the vectors X and Y.

A.3 LOW COHESION REFLECTIVE TEXT

Reading Block Divider: refLow_method e

One similarity between our tutored problems was that several of them used the same sequence of problem solving steps. Can you remember what they were?

The order of operations was often to look at the forces involved, net force, acceleration, velocity and displacement. In the Elevator problem we looked at all the forces involved and discovered there was only gravity. We then found the net force (gravity), the acceleration (downward), and finally the velocity and displacement (the same for man and keys).

The Car-truck question asked only about velocity. To answer it we reasoned from net force (same for car and truck) to acceleration (greater for car) to velocity (greater for car). It wasn't necessary to take the next step and reason about displacement.

In Earth-sun we applied the Third Law to show that the Sun's gravitational pull on the Earth was the same as the Earth's on the Sun. We did not need to think about net force, acceleration, velocity or displacement.

Notice that in no problem did it make sense to do steps out of order. We always had to look at all forces before finding the net force, for example. And we always had to find net force before finding acceleration. We couldn't have done them in the reverse order.

Reading Block Divider: refLow_vectors e

Vectors must be specified using both their magnitude and their direction, while scalars are specified using only their magnitude. Speed is a scalar quantity while velocity and acceleration are vector quantities.

The direction of a vector can be broken into independent horizontal and vertical components. Because they do not affect each other, they can be analyzed separately.

Do you remember which of the tutored problems took advantage of the decomposability of vectors?

The Plane-packet and the Pumpkin situations had motion in both the X and Y axes. We looked at the Y direction (which involved gravitational acceleration) separately from the X direction (in which there was no net force).

This approach was also used in the 'Car-truck' case. We were able to analyze the vertical direction (with no net force) separately from the horizontal direction (in which there was a net impact force).

In the first two problems mentioned, the vertical direction had a net force, and we were able to analyze it separately from the horizontal direction, which didn't. In the latter problem the horizontal had a net force, and we were able to ignore the vertical. In other problems there might be net forces in both the X and Y axes. Decomposability would still apply.

Reading Block Divider: refLow_firstLaw e

Newton's First Law can be stated as: 'Every body continues in its state of rest, or of motion in a straight line at constant speed, unless it is compelled to change that state by forces exerted upon it.' Do you remember which two problems used this law?

The First Law was useful in both the 'Plane-Packet' and 'Pumpkin' cases. In the first situation, before the load was dropped from the airplane they both had the same horizontal

velocity. After the release, there were no horizontal forces on the package (air-resistance was negligible). Therefore, by the First law, it would continue with a constant horizontal velocity.

In the 'Pumpkin' problem, it and the man had the same horizontal velocity before release. Afterwards, there were no horizontal forces on it (air-resistance was negligible). The pumpkin continued with a constant horizontal speed.

Other parts of these problems were quite different. The packet was dropped from the plane, without imparting any vertical force other than gravity. The pumpkin had both an initial upward velocity and an acceleration from gravity. In neither case did what was happening in the Y axis affect what was happening in the X axis. We were able to apply Newton's First Law, and realize that horizontal velocity would continue unchanged.

This law could also be applied in other situations. In the 'Elevator-Keys' problem, the man and keys were in a state of freefall, with only gravitational force acting on them. We concluded that both would have the same downward displacement. But why didn't they move sideways, relative to each other? Neither man nor keys had any horizontal forces acting on them, so both continued in their state of rest in the horizontal direction.

Reading Block Divider: refLow_thirdLaw e

In the Earth-Sun problem we had to compare the strength of the Sun's pull on the Earth with that of the Earth's on the Sun. In the Car-Truck problem we had to compare the force of the car's impact on the truck with that of the truck on the car. Do you remember which of Newton's Laws was useful in these cases?

In both situations we used the Third Law to show that the forces involved in an action/reaction pair had the same magnitude but acted in opposite directions. An action-reaction pair is formed whenever one object exerts a force on another object. Newton's Third Law says this force will have an equal and opposite reaction force. The type of force is always the same for both objects in the pair. It was gravitational on both Earth and Sun, and impact on both car and truck. They can operate along any axis, but always have opposite directions to each other. For example, the earth pulled in the opposite direction of the sun (vertically up vs vertically down), and the car's impact force was opposite to the truck's (horizontally right vs horizontally left).

In both of these situations, using Newton's Third Law allowed us to see that the forces acting on each object in the action/reaction pair had the same magnitude, even though the objects in the pair had different masses. The Earth pulls as hard on the Sun as the Sun does on it. The car hit the truck with as much force as the truck hit it, even though their masses were very different.

You can use the idea of an action/reaction pair to analyze any situation in which a force is exerted. After the plane releases it, the force of gravity becomes the net force, and the packet accelerates downward toward the Earth. Does the Earth accelerate toward the packet? Yes. Earth and packet form an action/reaction pair. The packet pulls on the Earth as hard as the earth pulls on it. The Earth therefore accelerates toward the packet, although less noticeably because of its greater mass.

Reading Block Divider: refLow_secondLaw e

Newton's Second Law was stated as 'The acceleration of a body is directly proportional to the magnitude of the net force acting on it and inversely proportional to its mass, and the acceleration is in the same direction as the net force.' 'Acceleration = net force / mass,' or 'net force = mass times acceleration': f = ma.

Newton's Second Law was used in almost all of our tutored problems. Can you think of how it applied in a few of them?

It was applied in essentially the same way in both the Pumpkin and Plane-packet situations. In both cases it was used to deduce acceleration from net force. In Plane-packet, the net force after release was gravity in the downward direction, so the package accelerated downward. In the Pumpkin problem, the net force was gravity, and the downward acceleration first reduced the upward velocity from the toss to zero, then increased it in the downward direction.

Newton's Second Law was used in the Car-truck question. We deduced (using the Third Law) that the horizontal impact force was equal for both objects. Using the formula f = ma, we deduced that because the car's mass was much less than that of the truck, the car's acceleration would be greater than the truck's.

F=ma is also useful in understanding an important part of the Elevator-keys problem. In the Car-truck problem both objects have the same impact force applied to them, but have

different masses and so gain different accelerations. Why then do the man and keys both have gravitational force applied to them, and have different masses, but gain the SAME acceleration? Gravity pulls with greater force on objects of greater mass. The greater mass increases an objects resistance to acceleration (acceleration is inversely proportional to mass). So, as mass increases both the force applied by gravity and the resistance to acceleration increase by the same amount. The net result is that freefall acceleration due to gravity is constant near the Earth's surface. This can be seen by writing Newton's Second Law as a = f/m. As an objects mass (m) increases, the force (f) applied to it (its weight) increases proportionally, and acceleration (a) remains the same.

Reading Block Divider: refLow_summary e

We have seen that certain key ideas occur frequently when solving force-and-motion problems. A fixed solution method allows us to break a problem into manageable parts and solve them in the correct order. Newton's First Law can be used to deduce the relationship between net force and acceleration. Newton's Second Law allows us to reason about the relationship between net force, mass and acceleration. Note that, similarly to the First Law, the Second Law can also be used to deduce the presence of an acceleration given a net force. Newton's Third law can be used to reason about the relative magnitude, direction and type of forces in an action-reaction pair.

Together, these ideas should be very useful in solving future force-and-motion problems.

A.4 HIGH COHESION REFLECTIVE TEXT

Reading Block Divider: refHigh_method e Problem Solving Method

One similarity between our tutored problems was that many of them used the same sequence of problem solving steps. Can you remember what those problem solving steps were?

The sequence of problem solving steps was to look first at the forces involved, then find the net force, then find the acceleration, then find the velocity and finally find the displacement. For example, in the Elevator problem we looked first at all the forces involved

and discovered there was only gravity. We then found the net force (gravity), the acceleration (downward), and finally the velocity and displacement (the same for man and keys), which solved the problem.

Not every step in the sequence is needed to answer every problem. The Car-truck problem, for example, asked only about velocity. To answer the Car-truck problem we reasoned from net force (same for car and truck) to acceleration (greater for car) to velocity (greater for car). It wasn't necessary to take the next step and reason about displacement.

The Earth-sun problem needed even fewer steps in the sequence because it asked only about the gravitational forces involved. We simply applied Newton's Third Law to show that the Sun's gravitational pull on the Earth was the same as the Earth's gravitational pull on the Sun. For this problem, we did not need to take the next steps in the sequence and find net force, acceleration, velocity or displacement.

Notice that in no problem did it make sense to do steps out of sequence. We always had to look at all forces before finding the net force, for example. And we always had to find net force before finding acceleration. We couldn't have done these steps in the reverse sequence.

Reading Block Divider: refHigh_vectors e Components of Vectors

Often an object's motion will be the result of several forces acting on it from different directions. In these cases it can be helpful to take advantage of the decomposability of vectors.

As you remember, vectors are quantities such as acceleration, which must be specified using both their magnitude and their direction. Scalar quantities, on the other hand, have no direction, and are specified using only their magnitude. This is how velocity differs from speed. Speed is a scalar quantity which is specified only by its magnitude. Velocity is a vector quantity which is specified using both its magnitude and its direction.

The direction of a vector quantity such as velocity is often easier to analyze if it is broken into separate horizontal and vertical components. The resulting horizontal and vertical components are independent. Because these components are independent, they can be analyzed separately.

Do you remember which of the tutored problems decomposed direction and analyzed the horizontal and vertical components separately?

The Plane-packet and the Pumpkin problems had motion in both the horizontal and vertical directions, but in both of those problems we were mainly interested in finding horizontal displacement. In both problems we decomposed the motion, and analyzed the vertical direction (which involved gravitational acceleration) separately from the horizontal direction (in which there was no net force).

We also decomposed motion in the 'Car-truck' problem. In the Car-Truck problem we were interested in horizontal velocity, and therefore in the horizontal forces. To arrive at horizontal velocity we analyzed the horizontal direction (in which there was a net impact force) separately from the vertical direction (in which there was no net force and no acceleration).

So, in the Plane-packet and Pumpkin problems the vertical direction had a net force, and we were able to analyze it separately from the horizontal direction, which didn't have a net force. In the Car-truck problem it was reversed, there was a net force in the horizontal direction, and we were able to ignore the vertical direction in which there was no net force. In other problems there might be net forces in both the horizontal and vertical directions, but the directions would still be decomposable. We would still be able to analyze them separately.

Reading Block Divider: refHigh_firstLaw e Newton's First Law

Newton's First Law can be stated as: 'Every body continues in its state of rest, or of motion in a straight line at constant speed, unless it is compelled to change that state by forces exerted upon it.' The First Law was useful in two of the problems we encountered during tutoring. Do you remember which two problems used Newton's First Law?

Newton's First Law was useful in both the 'Plane-Packet' and 'Pumpkin' problems. In the 'Plane-Packet' problem we wanted to determine if the packet had a horizontal displacement, and so we were interested in whether it had a horizontal velocity. Before the packet was dropped from the plane, both packet and plane had the same horizontal velocity. After it was dropped, there were no horizontal forces on the packet (remember that air-resistance was negligible). Newton's first Law told us that because there were no horizontal forces on the packet, it would continue with a constant horizontal velocity.

Similarly, in the 'Pumpkin' problem, before the man released the pumpkin, both man and pumpkin had the same horizontal velocity. After the man released the pumpkin, there

were no horizontal forces on it (air-resistance was again negligible), and so by Newton's First Law the pumpkin continued with a constant horizontal velocity.

Notice that the vertical aspects of these problems were quite different. The packet was dropped from the plane, without imparting any vertical force other than gravity. The pumpkin however was tossed, and so had both a vertically upward initial velocity and a vertically downward acceleration from gravity. However, because vectors are decomposable, in neither problem did what was happening in the vertical direction affect our analysis of the horizontal direction. Because of this, we were able to think about the horizontal direction separately from the vertical direction, apply Newton's First law, and realize that horizontal velocity would continue unchanged.

Newton's first Law could also be applied in other situations. For example, in the 'Elevator-Keys' problem, the man and keys were in a state of freefall, with only gravitational force acting on them. We concluded that the keys would have the same vertical displacement as the man. But why didn't the keys move horizontally, relative to the man? Well, neither man nor keys had any horizontal forces acting on them, so, as the First Law told us, both continued in their state of rest in the horizontal direction.

Reading Block Divider: refHigh_thirdLaw e Newton's Third Law

In the Car-truck problem we wanted to compare the relative accelerations of the car and truck. Therefore, we first had to compare the impact force of the car on the truck with the impact force of the truck on the car. Similarly, in the Earth-Sun problem we were asked to compare the force of the Sun's pull on the Earth with the force of the Earth's pull on the Sun. Do you remember which of Newton's Laws was useful in these two problems?

In these two problems we used Newton's Third Law to show that the forces involved in an action/reaction pair had the same magnitude but acted in opposite directions to each other. An action-reaction pair is formed whenever one object exerts a force on a second object. Newton's Third Law says that when one object exerts a force on a second object, there is an equal and opposite reaction force from the second object back onto the first object. In addition, the type of force is always the same for both objects in the action/reaction pair. For example it was gravitational force on both Earth and Sun, and impact force on both car and truck. The two forces in an action-reaction pair can operate along any axis, but

always have opposite directions to each other. For example, the earth pulled in the opposite direction than did the sun (vertically down vs vertically up), and the car's impact force was opposite to the truck's (horizontally right vs horizontally left).

In both the Car-truck and Earth-Sun problems, using Newton's Third Law allowed us to see that the forces acting on each object in the action/reaction pair had the same magnitude, even though the objects in the pair had different masses. The Earth pulls as hard on the Sun as the Sun pulls on it, even though the Sun is more massive. Similarly, the car hit the truck with as much force as the truck hit it, even though the truck had more mass.

You can use the idea of an action/reaction pair to analyze any problem in which one object exerts a force on another object. For example in the Plane-Packet problem, the Earth exerts a gravitational force on the packet, and the packet accelerates downward toward the Earth. Does the Earth also accelerate toward the packet? Yes. Earth and packet form an action/reaction pair, linked by gravitational attraction. The packet pulls on the Earth with gravity as hard as the earth pulls on it. The Earth therefore accelerates toward the packet, although less noticeably because of its greater mass.

Reading Block Divider: refHigh_secondLaw-a e Newton's Second Law

Newton's Second Law was stated as 'The acceleration of a body is directly proportional to the magnitude of the net force acting on it and inversely proportional to its mass, and the acceleration is in the same direction as the net force.' That is, acceleration = net force / mass. This is often rearranged and written as net force = mass times acceleration: f = ma.

Newton's Second Law was used in almost all of our tutored problems. Can you think of how it applied in a few of them?

Newton's Second Law was applied in essentially the same way in both the Pumpkin and Plane-packet problems. In both problems we first found the net force on an object. Newton's Second Law was then used to deduce that because there was a net force acting on it, that object would accelerate. In the Plane-packet problem, the net force after release was gravity in the downward direction, from which the Second Law allowed us to deduce that the package would accelerate in the downward direction. In the Pumpkin problem, the net force was again gravity, and the downward acceleration first reduced the upward velocity from the toss to zero, then increased it in the downward direction.

Newton's Second Law was also used in the Car-truck problem. In that problem, we first deduced (using Newton's Third Law) that the horizontal impact force was equal for both car and truck. Then, using the formula for Newton's Second Law, f = ma, we deduced that since the forces (f) were equal, but the car's mass (m) was much less than that of the truck, the car's acceleration (a) would be greater than the truck's.

Reading Block Divider: refHigh_secondLaw-b e

Newton's Second Law (F=ma) is also useful in understanding an important way in which the Elevator-keys problem differs from the Car-truck problem. In the Car-truck problem the car and truck have the same impact force applied to each, but they have different masses, and so gain different accelerations. But in the Elevator problem, man and keys both have gravitational force applied to them, and they also have different masses, but they gain the SAME acceleration. Why? The answer lies in realizing that gravity does not pull with equal force on objects of different mass. Gravity pulls with greater force on objects of greater mass. At the same time, however, the greater mass increases an objects resistance to acceleration (remember that acceleration is inversely proportional to mass). So, as mass increases both the force applied by gravity (its weight) and the resistance to acceleration increase by the same amount. The net result is that freefall acceleration due to gravity is constant near the Earth's surface. This can be seen by writing Newton's Second Law as a = f/m. As an objects mass (m) increases, the force (f) applied to it (it's weight) increases proportionally, and acceleration (a) remains the same.

Reading Block Divider: refHigh_summary e

Conclusion

We have seen that certain key ideas are frequently useful when solving force-and-motion problems. One key idea is our sequence of problem solving steps. Knowing a fixed sequence of problem solving steps allows us to break a problem into manageable parts and solve them in the correct order. Other useful ideas are Newton's three laws. Newton's First Law can be used to reason about the relationship between net force and acceleration. That is, when there is no net force, there will be no acceleration (or conversely, where there is no acceleration, there must be no net force). Newton's Second Law allows us to reason about the relationship between net force, mass and acceleration. For example, for a given force, an object with

less mass will accelerate more. Note that, similarly to the First Law, the Second Law can also be used to deduce the presence of an acceleration given a net force. Newton's Third law can be used to reason about the relative magnitude, direction and type of forces in an action-reaction pair.

Together, these ideas should be very useful in solving future force-and-motion problems.

APPENDIX B

COHMETRIX OUTPUT FOR LOW AND HIGH COHESION TEXTS

Measure Description	Low Cohesion	High Cohesion
CAUSVP 'Incidence of causal verbs, links, and particles'	54.53	56.38
CAUSC 'Ratio of causal particles to causal verbs'	0.48	0.68
CONADpi 'Incidence of positive additive connectives'	31.71	35.63
CONTPpi 'Incidence of positive temporal connectives'	8.24	9.92
CONCSpi 'Incidence of positive causal connectives'	15.22	20.75
CONADni 'Incidence of negative additive connectives'	5.07	6.77
CONTPni 'Incidence of negative temporal connectives'	0	0
CONCSni 'Incidence of negative causal connectives'	2.54	2.26
CONi 'Incidence of all connectives'	61.51	72.17
CREFA1u 'Argument Overlap, adjacent, unweighted'	0.42	0.56
CREFS1u 'Stem Overlap, adjacent, unweighted'	0.37	0.54
CREFP1u 'Anaphor reference, adjacent, unweighted'	0.24	0.21
CREFAau 'Argument Overlap, all distances, unweighted'	0.34	0.43
CREFSau 'Stem Overlap, all distances, unweighted'	0.32	0.42
CREFPau 'Anaphor reference, all distances, unweighted'	0.12	0.13
DENSNP 'Noun Phrase Incidence Score (per thousand words)'	278.38	274.7
DENSPR2 'Ratio of pronouns to noun phrases'	0.17	0.15
DENCONDi 'Number of conditional expressions, incidence score'	0	0.9
DENNEGi 'Number of negations, incidence score'	8.88	10.37
DENLOGi 'Logical operator incidence score'	38.05	40.14
LSAassa 'LSA, Sentence to Sentence, adjacent, mean'	0.29	0.39
LSApssa 'LSA, sentences, all combinations, mean'	0.3	0.39
LSAppa 'LSA, Paragraph to Paragraph, mean'	0.41	0.51
DENPRPi 'Personal pronoun incidence score'	46.93	41.5
HYNOUNaw 'Mean hypernym values of nouns'	4.83	4.91
HYVERBaw 'Mean hypernym values of verbs'	1.29	1.34
READNP 'Number of Paragraphs'	51	70
READNS 'Number of Sentences'	116	152
READNW 'Number of Words'	1577	2217

Table 41: Complete CohMetrix output for high and low cohesion texts, Part 1

Measure Description	Low Cohesion	High Cohesion
READAPL 'Average Sentences per Paragraph'	2.28	2.17
READASL 'Average Words per Sentence'	13.6	14.59
READASW 'Average Syllables per Word'	1.56	1.58
READFRE 'Flesch Reading Ease Score (0-100)'	61.15	58.28
READFKGL 'Flesch-Kincaid Grade Level (0-12)'	8.11	8.75
SYNNP 'Mean number of modifiers per noun-phrase'	1.02	1.01
SYNHw 'Mean number of higher level constituents per word'	0.71	0.7
SYNLE 'Mean number of words before the main verb of main clause'	3.03	3.61
TYPTOKc 'Type-token ratio for all content words'	0.34	0.27
FRQCRacw 'Celex, raw, mean for content words'	2445.22	2752.88
FRQCLacw 'Celex, logarithm, mean for content words'	2.17	2.19
FRQCRmcs 'Celex, raw, minimum in sentence for content words'	40.84	59.01
FRQCLmcs 'Celex, log., minimum in sentence for content words'	1.14	1.1
WORDCacw 'Concreteness, mean for content words'	366.38	360.66
CONLGpi 'Incidence of positive logical connectives'	18.39	26.61
CONLGni 'Incidence of negative logical connectives'	8.24	9.47
INTEC 'Ratio of intentional particles to intentional content'	0	0
INTEi 'Incidence of intentional actions, events, and particles.'	12.05	9.92
TEMPta 'Mean of tense and aspect repetition scores'	0.87	0.86
STRUTa 'Sentence syntax similarity, adjacent'	0.08	0.08
STRUTt 'Sentence syntax similarity, all, across paragraphs'	0.08	0.08
STRUTp 'Sentence syntax similarity, sentence all, within paragraphs'	0.09	0.08
CREFC1u 'Prop. of content words that overlap between adj. sent.'	0.11	0.16
SPATC 'Mean of location and motion ratio scores.'	0.52	0.49
WORDCmcs 'Concreteness, minimum in sentence for content words'	158	158
GNRPure 'Genre purity'	0.5	0.5
TOPSENr 'Topic sentence-hood'	0.18	0.21

Table 42: Complete CohMetrix output for high and low cohesion texts, Part 2

APPENDIX C

TEST QUESTIONS

post-mc1

A girl riding a bike in a straight line at a constant speed drops an ice cream cone she was holding. Immediately after she drops the ice cream cone, what is the relationship between the horizontal speed of the girl and the horizontal speed of the cone? (Assume air resistance is negligible)

- the horizontal speed of the girl is greater than the horizontal speed of the cone
- the horizontal speed of the cone is greater than the horizontal speed of the girl
- the horizontal speed of the girl is the same as the horizontal speed of the cone CORRECT
- there is not enough information to answer

post-mc2 B33

A frustrated programmer throws her laptop out of the window of a tall building with an initial velocity, Vi, in the horizontal direction. Assuming air resistance is negligible, what is true of the horizontal component of velocity of the laptop while it is falling?

- it will increase
- it will decrease
- it will remain the same CORRECT
- there is not enough information to answer

post-mc3 B34

Suppose that a rollerblader is skating down a city street and maintains a constant horizontal velocity. You pull alongside the rollerblader in your car. Just then you get a call on you cell phone. You maintain you speed, ignoring the rollerblader. At the end of your call, you look out the window. What should you see?

- The rollerblader has pulled ahead of you
- The rollerblader is skating along beside you CORRECT
- the rollerblader has fallen behind

- Other:

post-mc4 B13

A woman carrying her groceries home is walking north at 1 m/s. A jogger is moving northeast; the northern component of his velocity is 1 m/s. How does the norther displacement of the jogger compare to the norther displacement of the woman at any time?

- they are the same CORRECT
- the northern displacement of the jogger is greater than that of the woman
- the northern displacement of the woman is greater than that of the jogger
- there is not enough information to answer

post-mc5 18Q(far)

A group of boys, including Albert and Bill, are playing a game of it-tag. Albert, who has just been tagged 'it', runs 40 meters north, then 20 meters east, and 10 meters south, where he arrives at t1. At the time Albert was tagged, Bill was 10 meters north of Albert and runs 20 meters north, then 20 meters east, where he arrives at t1. Does Albert catch Bill?

- Yes CORRECT
- No
- there is not enough information to answer

post-mc6 B14

A model rocket is launched vertically. When it is 1,000 meters above the ground it loses all engine power and breaks into two pieces, the front end and the rear end, which has the failed engine. After the engine power is lost which of the two pieces of the rocket is/are in free fall?

- the front end
- the rear end
- both ends CORRECT
- neither end
- Other:

post-mc7 B15

A stuntwoman drives a motorcycle up a ramp and jumps over a row of cars and lands safely, remaining on her bike. When it lands, the motorcycle is moving at 20 m/s. What is the speed of the stuntwoman when she lands?

- 20 m/s CORRECT
- a little faster than 20m/s
- a little slower than 20 m/s
- there is not enough information
- Other:

post-mc8 18Q(far)

A stuntwoman drives a motorcycle up a ramp and jumps over a row of cars and lands safely, remaining on her bike. When it lands, the motorcycle is moving at 20 m/s. After her landing, the stuntwoman decelerates at a rate fo 5 m/s^2. If the stuntwoman remains on her motorcycle while decelerating, what is the magnitude of the deceleration of the motorcycle?

- 5 m/s^2 CORRECT
- less than 5 m/s^2
- greater than 5 m/s^2
- there is not enough information to answer this question

post-mc9B18

An old Volkswagon has a maximum acceleration of 2 m/s^2 as it starts from a stop sign; a cyclist can also accelerate at a maximum rate of 2 m/s^2. Starting at rest, the Volkswagon and cyclist accelerate at their maximum rates for the same time. Which has traveled farther in that time?

- the Volkswagon
- the cyclist
- they travel the same distance CORRECT
- there is not enough information to answer

post-mc10B19

An airplane is taking off from the ground. In which direction does gravity act on the airplane while it is taking off?

- vertically down, or toward the center of the earth CORRECT
- initially horizontally, then increasingly vertical
- almost vertically down, but slightly angled due to the rotation of the earth
- in the direction of the motion of the airplane
- Other:

post-mc11 18Q(far)

Two friends, Thelma and Louise, are roommates who share a car and also work at the same store. Thelma plans to drive the car along a straight highway from her house to the store, where Louise will then drive the car home. Thelma is late for her shift, and accelerates at a constant rate of 0.1 m/s^2 from her house to the store. Later on, Louise, who is tired and wants to get home to sleep, also accelerates at a constant rate of 0.1 m/s^2 from the store to her home taking the same straight highway that Thelma took. What is the relationship between the time it took Thelma to drive to the store and the time it took Louise to drive from the store to her home?

- it took Thelma longer
- it took Louise longer
- it took them the same time CORRECT
- there is not enough information to answer

post-mc12 B21

Two rugby players, Collin and Ewan, of the same mass, are running toward each other and collide with each other. The magnitude of the force of Collin on Ewan is 500 N. What is the magnitude of the force of Ewan on Collin?

- 500 N CORRECT
- less then 500 N
- more information is necessary to answer this question

post-mc13 B22

Two rugby players, Collin and Ewan, of the same mass, are running toward each other and collide with each other. The magnitude of the force of Collin on Ewan is 500 N. If the direction of the force Collin exerts on Ewan is to the east, what is the direction of the force that Ewan exerts on Collin?

- Also to the east
- to the west CORRECT
- more information is necessary to answer this question

post-mc14 18Q(far)

A truck driver applies the brakes to stop for a red light. The magnitude of the acceleration of the truck while stopping is 2 m/s^2. The driver is wearing his seatbelt and does not move relative to his seat while he slows. What is the magnitude of the acceleration of the driver while the truck comes to a stop?

- 2 m/s^2 CORRECT
- less than 2 m/s^2
- greater than 2 m/s^2
- there is not enough information to answer
- Other:

post-mc15 18Q(far)

A truck driver applies the brakes to stop for a red light. The magnitude of the acceleration of the truck while stopping is 2 m/s^2. The driver is wearing his seatbelt and does not move relative to his seat while he slows. During the period when the truck slows, a contact force is acting on the truck driver. What object exerts this contact force?

- the tires of the truck
- the seat of the truck
- the seatbelt CORRECT
- the road
- Other:

post-mc16 18Q(far)

A man puts his small child on his shoulders and begins walking horizontally. Is it necessary for the man to apply a horizontal force on his child if the child is to move with her father when he starts walking?

- Yes, the child will fall forward if he does not
- Yes, the child will fall backwards if he does not CORRECT
- no force is necessary . The child will move with her father
- Other:

post-mc17 18Q(far)

The American bobsledders have a bobsled that weighs 300 N. The Jamaican's bobsled weighs 280 N. If the bobsledding teams are equally strong (i.e., they can push on the bobsled with the same force) and push the sled for the same period of time until they jump into it, which team's bobsled will be moving faster at the moment they jump into it?

- The American team's
- The Jamaican team's CORRECT
- They will be the same
- there is not enough information to answer

post-mc18 18Q(far)

On the earth, a coin is dropped and accelerates at a rate, A, until it hits the ground a short time, t, later. An astronaut on the moon drops another similar coin, which accelerates at a rate less than A until it hits the moon's surface in the same time, t. What is the relationship between the magnitude of the velocities of the coin dropped on the earth and the coin dropped on the moon right before they land?

- the final velocity of the coin dropped on the moon is greater
- the final velocity of the coin dropped on the earth is greater CORRECT
- the final velocities are the same
- there is not enough information to answer

post-mc19 18Q(far)

A motorboat's engine is running at it's maximum force when it's driver accidentally drops his 100-kg cooler overboard. When the cooler falls overboard what happens to the force produced by the engine?

- it decreases
- it increases
- it remains the same CORRECT
- there is not enough information to answer

post-mc20 **post-mc20 18Q(far)**

A 500-kg horse is walking along carrying a 100-kg woman when the horse is startled by a snake and begins to run. The force of the ground on the horse's hoofs as the horse speeds up is 1200 N. If this is the only force acting on the horse in the horizontal direction while it is taking off, what is the horizontal acceleration of the horse?

- 12 m/s^2
- 2.4 m/s^2

- 2 m/s^2 CORRECT
- Other:

post-mc21 B23

A wild horse runs at full speed, and jumps over the fence that had been trapping him. While he is in the air during his jump, in which direction is his acceleration due to the earth's gravitation force?

- vertically downward CORRECT
- both vertically downward and horizontally in the direction the horse is moving
- only horizontally in the direction the horse is moving
- he is not accelerating due to the earth's gravity while the horse is in the air

post-mc22B24

An empty chair lift of mass, mcl, is moving up a hill at a 45-degree angle with a constant speed of 2 m/s. What is the magnitude of the acceleration of the chair lift?

- 0 m/s^2 CORRECT
- 2 m/s^2
- 9.8 m/s^2
- 11.8 m/s^2
- Other:

post-mc23 B17

An empty chair lift of mass, mcl, is moving up a hill at a 45-degree angle with a constant speed of 2 m/s. What is the magnitude of the net force acting on the chair-lift while it is moving up the hill?

- 0 Newtons CORRECT
- mccl * g (where g is the acceleration due to gravity) Newtons
- 2 * mb Newtons
- 11.8 * mb Newtons
- Other:

post-mc24 B35

A stuntman jumps horizontally off the edge of a tall building and freefalls until he lands safely on the foam mats below. At the moment the stuntman stps off the building, a stunt dog also jumps from the edge of the same building (and, being a professional, also lands safely). Which of the following is true?

- the stuntman will land first
- the stunt dog will land first
- the man and the dog will land at the same time CORRECT
- there is not enough information to answer the question

post-mc25 B11

Using a slingshot, you shoot a stone straight up into the air. Does the force from the slingshot continue to act on the stone after the stone leaves the slingshot?

- Yes. The stone will still feel the force from the slingshot.
- No. The stone will only feel the force from the slingshot while the slingshot is pushing on it. CORRECT
- Yes, but only for a short while. After a few seconds the force from the slingshot has dissipated.

post-mc26 B25

Using a slingshot, you shoot a stone straight up into the air. When the stone is halfway between where it was first shot and it's maximum height (and is still moving upward), which force(s) act on it?

- gravity CORRECT
- the force of the slingshot
- there are no forces acting on it
- both gravity and the force of the throw

post-mc27 B29

Using a slingshot, you shoot a stone straight up into the air. When the stone shot from the slingshot reaches it's maximum height, which force(s) are acting on it?

- gravity CORRECT
- the force of the slingshot
- there are no forces acting on it at that instant
- both gravity and the force of the slingshot

post-mc28 B26

Assume you have two balloons: one is filled with water and the other is filled with oil. The balloon filled with oil weighs less than the balloon filled with water, but they are identically shaped. You drop both balloons from a bridge at the same time. What is the relationship between the accelerations of the heavier (water-filled) and the lighter (oil-filled) balloons as they fall to the river below?

- the acceleration of the heavier balloon is the same as that of the lighter balloon CORRECT
- the acceleration of the heavier balloon is greater than that of the lighter balloon
- the acceleration of the heavier balloon is less than that of the lighter balloon

post-mc29B27

A block slides in a straight line across a flat frictionless surface at a speed of 2 m/s. When the block is moving due east, a strong force begins to push on the block to the north. Which is true of the speed of the block in the eastern direction several seconds after the force begins to act on it?

- it is greater than 2 m/s

- it is less than 2 m/s
- it is 2 m/s CORRECT
- insufficient information to tell

post-mc30 18Q(far)

A shot putter throws the shot (a very heavy metal ball) at a 45-degree angle above the horizon. As the shot flies through the air, which of the following forces will significantly affect it's motion?

- gravitational force CORRECT
- the force from the shot putter's throw
- air resistance

post-mc31B28

a 100 kg man is skating clockwise around a skating rink at 3 m/s. A small 30 kg child is skating counter-clockwise (in the opposite direction from everyone else) around the skating rink at a speed of 2 m/s. Neither are paying attention and the child runs into the man. The magnitude of the force the man exerts on the child is 30 N. What is the magnitude of the force the child exerts on the man?

- 30 N CORRECT
- less than 30 N
- 0 N (the child does not exert a force on the man)
- greater than 30 N (because the child hit hte man, he exerts a greater force on the man than the man exerts on him)

post-mc32 B16

a 100 kg man is skating clockwise around a skating rink at 3 m/s. A small 30 kg child is skating counter-clockwise (in the opposite direction from everyone else) around the skating rink at a speed of 2 m/s. Neither are paying attention and the child runs into the man. The magnitude of the force the man exerts on the child is 30 N. During the collision between the man and child, which is true of the relationship between the magnitudes of the acceleration of the man and the acceleration of the child?

- the acceleration of the man is greater than the acceleration of the child
- the acceleration of the child is greater than the acceleration of the man CORRECT
- the accelerations of the man and child are equal
- there is not enough information

post-mc33 18Q(far)

A hot air balloon carrying sightseers is moving downward and is slowing down for a landing. What can you say about the sum of the forces acting on the balloon in the vertical direction?

- there are no forces acting on the balloon in the vertical direction
- the sum of the forces act downward CORRECT

- the sum of the forces on the balloon act upward

post-mc34B30 **post-mc34 B30**

A cross-country skier is skiing horizontally across the snow and before realizing it, has skied onto a large patch of ice. The ice is very slippery and her skis are waxed, so that there is no frictional force between the skis and ice, and there are no other forces acting on the skier in the horizontal direction. What happens to the horizontal speed of the skier while she is on the ice patch?

- because there are no forces acting on the skier in the horizontal direction, the skier's horizontal speed decreases
- because there are no forces acting on the skier in the horizontal direction, the skier's horizontal speed remains constant CORRECT

post-mc35B20

A research satellite orbits the earth at a fixed distance of 100 miles from the earth's surface. Does the earth's gravity act on the satellite?

- Yes. The earth's gravity acts on everything near it's surface CORRECT
- Because the satellite is orbiting the earth, it is accelerating toward the earth; thus the earth does not exert a gravitational force on the satellite
- Because the satellite is so high above the earth's surface, it does not experience the earth's gravitational force

post-mc36B36W

A distant planet has a moon that orbits around it. Which of the following statements is true:

- The planet exerts a gravitational force on the moon, but the moon does not exert a gravittational force on the planet
- Both the planet and the moon exert a gravitational force on the other, but the force of the planet on the moon is greater than the force of the moon on the planet
- Both the planet and the moon exert a gravitational force on the other, and the gravitational force of the moon on the planet is the same as the force of the planet on the moon CORRECT

post-mc37 B31

In the filming of Mission Impossible 4, Tom Cruise insists on performing his own stunt, which involves riding a 200-kg motorcycle straight into a 100,000-kg tractor-trailer traveling in the opposite direction, then jumping off a moment before the motorcycle and tractor-trailer collide. The motorcycle is smashed to half it's length during the collision, whereas the tractor-trailer's front end is merely dented in a few inches. Which of the following is true of the relationship between the force of the motorcycle and the tractor-trailer and the force of the tractor-trailer on the motorcycle?

- Because the motorcycles's acceleration during the collision was greater than the tractor-trailer's acceleration, the force of the motorcycle on the tractor-trailer is greater than the force of the tractor-trailer on the motorcycle
- the force of the tractor-trailer on the motorcycle is greater than the force of the motorcycle on the tractor-trailer
- the force of the tractor-trailer on the motorcycle is equal to the force of the motorcycle on the tractor-trailer CORRECT

post-mc38 18Q(far)

Which of the following statements is true about the relationship between the sun and the earth's motion?

- the earth orbits the sun, but the sun does not orbit the earth CORRECT
- the sun orbits the earth, but the earth does not orbit the sun
- the sun and the earth orbit each other
- neither the sun nor the earth orbit the other

post-mc39B32

You are sitting on a subway train facing the wrong way (facing the back of the train, which is west) when the train accelerates east and the upper half of your body lurches west (in the direction you are facing) while you remain on your seat. While you lurch west, which of the following is true:

- there is a westward force on you, in the same direction you lurch
- there is an eastward force on you (opposite the direction you lurch) CORRECT
- there are no forces acting on you

post-mc40 18Q(far)

A basketball player bounces a basketball on the floor of a gymnasium. Which of the following is true of the relationship between the force of the ball on the gym floor and the force of the gym floor on the ball when the ball hits the floor?

- the ball exerts a force on the floor, but because it does not move, the floor does not exert a force on the ball
- the floor exerts a force on the ball, but the ball does not exert a force on the floor
- the ball and floor both exert a force on the other CORRECT

post-sled1 18Q(far)

Two identical twins are sitting in identical sleds, which are resting on friction-free ice. Behind each twin on the sled is a pile of bricks. Each brick weighs 1kg. In the first sled, the first twin has a pile of four bricks. In the second sled, the second twin has a pile of three bricks. At exactly the same time each twin begins picking bricks off her pile and throwing them horizontally behind her. The twins are identical, so they weigh the same and throw with exactly the same amount of force. They also throw at the same rate, one brick per second.

What happens to the sleds, when the twins begin throwing bricks?

- Nothing, both remain stationary.
- They both move
- Only the first sled moves
- Only the second sled moves

post-sled2 18Q(far)

Two identical twins are sitting in identical sleds, which are resting on friction-free ice. Behind each twin on the sled is a pile of bricks. Each brick weighs 1kg. In the first sled, the first twin has a pile of four bricks. In the second sled, the second twin has a pile of three bricks. At exactly the same time each twin begins picking bricks off her pile and throwing them horizontally behind her. The twins are identical, so they weigh the same and throw with exactly the same amount of force. They also throw at the same rate, one brick per second.

When the first brick is being thrown from each sled, which sled has the greater acceleration?

- Neither, both have identical acceleration.
- The first sled.
- The second sled.

post-sled 3 18Q(far)

Two identical twins are sitting in identical sleds, which are resting on friction-free ice. Behind each twin on the sled is a pile of bricks. Each brick weighs 1kg. In the first sled, the first twin has a pile of four bricks. In the second sled, the second twin has a pile of three bricks. At exactly the same time each twin begins picking bricks off her pile and throwing them horizontally behind her. The twins are identical, so they weigh the same and throw with exactly the same amount of force. They also throw at the same rate, one brick per second.

After the last brick has been thrown, which sled has the greater velocity?

- Neither, both have identical velocity
- The first sled
- The second sled

cube1 18Q(far)

A metal cube is suspended by a spring near an electro-magnet. When a switch is thrown, current flows through the magnet, creating a magnetic field which pulls the cube downward, stretching out the spring.

What is the reaction force to this pull?

- The reverse pull of the spring
- Electrical resistance in the magnet
- The cube's magnetic pull on the magnet.

APPENDIX D

MCNAMARA'S TEXTS

D.1 HEART DISEASE HIGH COHESION

D.1.1 Heart Disease

The heart is the hardest-working organ in the body. We rely on it to supply blood regularly to the body every moment of every day. Any disorder that stops the heart from supplying blood to the body is a threat to life. Heart disease is such a disorder. It is very common. More people are killed every year in the U.S. by heart disease than by any other disease.

There are many kinds of heart disease, some of which are present at birth and some of which are acquired later.

D.1.1.1 1. Congenital heart disease A congenital heart disease is a defect that a baby is born with. Most babies are born with perfect hearts. But one in every 200 babies is born with a bad heart. For example, hearts have flaps, called valves, that control the blood flow between its chambers. Sometimes a valve develops the wrong shape. It may be too tight, or fail to close properly, resulting in congenital heart disease. Sometimes a gap is left in the wall, or septum, between the two sides of the heart. This congenital heart disease is often called a "septal defect". When a baby's heart is badly shaped, it cannot work efficiently. It cannot pump enough blood through the lungs so that it receives enough oxygen. As a result, the baby becomes breathless. The blood also cannot get rid of carbon

dioxide through the lungs. Therefore, the blood becomes purplish, which causes the baby's skin to look blue. Fortunately, it is now possible to save the lives of many "blue babies".

D.1.1.2 2. Acquired heart disease Some heart diseases are acquired after the baby is born. Rheumatic fever is an example of an acquired disease that may cause damage to the heart. This disease usually follows a sore throat caused by bacteria known as streptococci. This is often called "strep throat". When strep throat causes rheumatic fever, the tissues of the heart become inflamed. If the heart is badly affected, it fails very soon. Usually, however, it recovers, and the results of the damage are seen only years later. This is because the rheumatic fever leaves scars in the valves of the heart. Therefore, they cannot work properly. This puts a strain on the heart so that eventually it may fail. The effects of the rheumatic fever may take up to twenty or thirty years to appear.

Coronary disease is another example of an acquired heart disease. This disease affects the coronary arteries. These are the blood vessels that extend across the heart and supply it with blood from the lungs. They are very important because they give the heart muscle the oxygen it needs to carry on working. In coronary disease the coronary arteries become blocked, causing parts of the heart muscle to die because of the lack of oxygen. When this happens, the patient has a heart attack, which can be fatal. The blockage of a coronary artery is usually caused by a clot of blood, called a "thrombus". When a clot forms in a coronary artery, this is called "coronary thrombosis." That is the correct name for a heart attack. In normal arteries, blood does not form clots. But in coronary disease, the walls of the arteries are not normal. They become lumpy, rough, and narrow. The lumps break off and form clots that stop the flow of blood to the heart.

Other examples of acquired heart disease are arrhythmia, angina, and high blood pressure. Arrhythmia, which means "lack of rhythm", is an interruption of the heart's normal beat. Angina is a sharp pain in the chest which is very similar to that caused by a heart attack, or thrombosis. High blood pressure is one of the most common heart diseases. It places a heavy strain on the heart and other organs. Therefore, if it is not treated, high blood pressure may lead to heart attacks, kidney failure, or other serious problems. High blood pressure is a disease which has no symptoms. Thus, a person may not be aware of

having it unless the blood pressure is measured.

D.1.1.3 3. Treatment and prevention of heart disease Since the mid-1960's, medical science has made tremendous progress in the treatment and prevention of heart disease. Both new drugs and new surgical methods have been developed. Among the new drugs for treating heart disease are chemicals called "beta blockers". The beta-blockers lessen the after-effects of heart attacks; they can prevent second attacks; and they can lower the blood pressure of people who have high blood pressure. Other drugs dissolve the lumps which break off the walls of arteries so that they do not stop the flow of blood to the heart.

Surgical techniques for treating heart disease range from repairing or replacing damaged parts, such as valves or arteries, to replacement of the entire heart. If a heart has been so damaged that it can no longer function, it can be replaced by a mechanical heart, or, more often, by a heart transplant. In transplant surgery, the healthy heart of someone who has died replaces the diseased heart of the patient. Mechanical devices can be implanted in people's bodies to keep their hearts functioning. The pacemaker is the most common of these devices. It does not heal the diseased heart, but it relieves the symptoms of an irregular heart beat and maintains the steady beat needed for normal living. When a heart cannot pump enough blood through the lungs because of poorly functioning valves, the valves can be replaced with artificial ones of plastic and metal. For patients with coronary disease, "by-pass surgery" is often used to repair clogged or damaged arteries. Doctors use pieces of a patient's own veins, often from the leg, to replace the damaged portions of arteries.

Preventive care is also getting better as scientists learn more and more about the causes of heart disease. They have shown that diet can be an important means of controlling heart disease. For example, a substance called cholesterol is known to cause a build-up of fatty substances in the blood vessels, which can cause blood clots to form in the arteries. Therefore, doctors stress the importance of a diet low in cholesterol. Similarly, salt is known to increase blood pressure, so doctors recommend a low-salt diet for patients with high blood pressure.

D.2 HEART DISEASE LOW COHESION

D.2.1 Heart Disease

The heart is the hardest-working organ in the body. We rely on a regular blood supply every moment of every day. Any disorder that stops the blood supply is a threat to life. Heart disease is very common. More people are killed every year in the U.S. by heart disease than by any other disease.

A congenital disease is one that a person is born with. Most babies are born with perfect hearts. In about one in every 200 cases something goes wrong. Sometimes a valve develops the wrong shape. It may be too tight, or fail to close properly. Sometimes a gap is left in the septal wall between the two sides of the heart. This is often called a septal defect. When a baby's heart is badly formed, it cannot work efficiently. The blood does not receive enough oxygen. The baby becomes breathless. The blood cannot get rid of carbon dioxide through the lungs. It becomes purplish, and the baby's skin looks blue. It is now possible to save the lives of many blue babies.

The disease called rheumatic fever may cause harm to the heart. The disease usually follows a sore throat caused by bacteria called streptococci. The tissues of the heart become inflamed. If it is badly affected, it fails. Usually it recovers, and the results of the damage are seen only years later. The valves of the heart are left with scars. They cannot work properly. This puts a strain on the heart. Eventually it may fail. The effects of the rheumatic fever may take up to twenty or thirty years to appear.

The blood vessels that extend across the heart and supply it with blood are called the coronary arteries. They are very important. They give the heart the oxygen it needs to carry on working. If they become blocked, parts of the heart muscle will die. The patient has a heart attack, which can be fatal. The blockage of a coronary artery is usually caused by a thrombus, or blood clot. Coronary thrombosis happens when a clot forms in a coronary artery. That is the correct name for a heart attack. In normal arteries, blood does not form clots. In coronary disease, the walls of the blood vessels become lumpy, rough, and narrow.

Arrhythmia is an interruption of the heart's normal beat Angina is a sharp pain in

the chest which is very similar to that caused by thrombosis. High blood pressure is very common. If untreated, high blood pressure may lead to heart attacks, kidney failure, or other serious problems. High blood pressure may have no symptoms. A person may not be aware of having it unless the blood pressure is measured.

Among the new drugs for treating heart disease are a family of compounds called beta blocking drugs, or simply, beta blockers. They lessen the after-effects of heart attacks, can prevent second attacks, and can lower the blood pressure of people who have high blood pressure. Other drugs dissolve the lumps which break off the walls of veins and arteries.

Heart transplants are used more often than mechanical hearts. In transplant surgery, the healthy heart of someone who has died replaces the heart of the patient. Mechanical devices can be implanted in people's bodies to keep their hearts functioning. The commonly used pacemaker does not heal the diseased heart, but it relieves the symptoms of an irregular heart and keeps a steady beat for normal living. When a heart cannot pump enough blood through the lungs because of poorly functioning valves, the valves can be replaced with artificial ones of plastic and metal. By-pass surgery is used to repair clogged or damaged blood vessels. Doctors use pieces of a patient's own veins, often from the leg, to replace the damaged portions of arteries.

A substance called cholesterol is known to cause a build-up of fatty substances in the blood vessels, which can lead to heart disease, so doctors stress the importance of a diet low in fats. Salt is known to increase the blood pressure, so a low-salt diet is recommended.

APPENDIX E

BIBLIOGRAPHY

Gregory Aist, Barry Kort, Rob Reily, Jack Mostow, and Rosalind Picard. Experimentally augmenting an intelligent tutoring system with human-supplied capabilities: Adding human-provided emotional scaffolding to an automated reading tutor that listens. *Proceedings of the Fourth IEEE International Conference on Multimodal Interfaces*, 71(2):126–148, 2003.

V. Aleven and K.R. Koedinger. An effective metacognitive strategy: Learning by doing and explaining with a computer-based cognitive tutor. *Cognitive Science*, 26:147 – 179, 2002.

Patricia Alexander, Jonna Kulikowich, and Sharon Schulze. The influence of topic knowledge, domain knowledge, and interest on the comprehension of scientific exposition. *Learning and Individual Differences*, 6(4):379–397, 1994.

Robert K. Atkinson, Mary M. Merrill, and Alexander Renkl. Transitioning from studying examples to solving problems: Effects of self-explanation prompts and fading worked-out steps. *Journal of Educational Psychology*, 95:774 – 783, 2003.

Regina Barzilay and Mirella Lapata. Modeling local coherence: An entity-based approach. In *Proceedings of the 43rd Annual Meeting of the ACL*, pages 141–148, Ann Arbor, June 2005.

Philip Bell and Elizabeth Davis. Designing mildred: Scaffolding students' reflection and argumentation using a cognitive software guide. *Fourth International Conference of the Learning Sciences*, pages 142 – 149, 2000.

B. S. Bloom. The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher*, 13:4 – 16, 1984.

Pietro Boscolo and Lucia Mason. Topic knowledge, text coherence, and interest: How they interact in learning from instructional texts. *The Journal of Experimental Education*, 71(2):126–148, 2003.

Bruce K. Britton and Sami Gulgoz. Using kintsch's computational model to improve instructional text: Effects of repairing inference calls on recall and cognitive structures. *Journal of Educational Psychology*, 83:329 – 345, 1991.

Min Chi, Kurt Vanlehn, and Diane Litman. The more the merrier? examining three interaction hypotheses. *Proceedings of the 32nd Annual Conference of the Cognitive Science Society (CogSci2010)*, 2010.

M.T.H. Chi, R. Glaser, and E. Rees. Expertise in problem solving. In R. Sternberg, editor, *Advances in the Psychology of Human Intelligence*, pages 7 – 76. Erlbaum, Hillsdale, 1982.

Laurent Cimolino, Judy Kay, and Amanda Miller. Incremental student modelling and reflection by verified concept-mapping. *Supplementary Proceedings of the AIED2003: Learner Modelling for Reflection Workshop*, 2003.

Patricia Cohen, Jacob Cohen, Leona Aiken, and Stephen West. The problem of units and the circumstance for pomp. *Multivariate Behavioral Research*, 34(3):315–346, 1999.

Allan Collins and John Seely Brown. The computer as a tool for learning through reflection - technical report. (376), 1986.

John Connelly and Sandra Katz. Toward more robust learning of physics via reflective dialogue extensions. In G. Siemens and C. Fulford, editors, *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications*, pages 1946–1951, Chesapeake, VA, 2009. IOS Press.

Albert Corbett, Angela Wagner, Sharon Lesgold, Harry Ulrich, and Scott Stevens. The impact on learning of generating vs. selecting descriptions in analyzing algebra example solutions. *Proceedings of the 7th International Conference on Learning Science*, pages 99 – 105, 2006.

Mark Core, Johanna Moore, and Claus Zinn. The role of initiative in tutorial dialogue. In *10th Conference of the European Chapter of the Association for Computational Linguistics*, Budapest, Hungary, April 2003.

Jose Cortina. What is coefficient alpha? an examination of theory and applications. *Journal of Applied Psychology*, 78(1):98–104, 1993.

Lee Cronbach. Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3):297–334, 1951.

Elizabeth Davis. Prompting middle school science students for productive reflection: Generic and directed prompts. *The Journal of the Learning Sciences*, 2:91 – 142, 2003.

Elizabeth Davis and Marcia Linn. Scaffolding students' knowledge integration: prompts for reflection in kie. *International Journal of Science Education*, 22:819 –837, 2000.

Teresa del Soldato and Benedict du Boulay. Implementation of motivational tactics in tutoring systems. *Journal of Artificial Intelligence in Education*, 6(4):337–378, 1995. ISSN 1043-1020.

John Dewey. *How We Think*. 1910.

S. D'Mello, P. Hays, C. Williams, W. Cade, J. Brown, and A. Olney. Collaborative lecturing by human and computer tutors. In J. Kay and V. Aleven, editors, *10th International Conference on Intelligent Tutoring Systems*, pages 609–618. Springer:Berlin / Heidelberg., 2010.

Myroslava Dzikovska, GwendolynCampbell, Charles Callaway, NatalieSteinhauser, Elaine Farrow, Johanna Moore, Leslie Butler, and Colin Matheson. Diagnosing natural language answers to support adaptive tutoring, 2008.

Sidney K. Dmello, Scotty D. Craig, Barry Gholson, and Stan Franklin. Integrating affect sensors in an intelligent tutoring system. In *Affective Interactions: The Computer in the Affective Loop Workshop at 2005 Intl. Conf. on Intelligent User Interfaces*, pages 7–13. AMC Press, 2005.

Leonard Feldt. The use of extreme groups to test for the presence of a relationship. *Psychometrika*, 26(3):307–316, 1961.

P.W. Foltz, W. Kintsch, and T.K. Landauer. The measurement of textual coherence with latent semantic analysis. *Discourse Processes*, 25:285 – 308, 1998.

Kate Forbes-Riley and Diane Litman. Designing and evaluating a wizarded uncertainty-adaptive spoken dialogue tutoring system. In *Computer Speech and Language*, In press.

John Frederiksen and Barbara White. Cognitive facilitation: a method for promoting reflective collaboration. *Proceedings of the 2nd international conference on Computer support for collaborative learning*, pages 55–64, 1997.

Reva Freedman. Using tutoring patterns to generate more cohesive text in an intelligent tutoring system. In *ICLS '96: Proceedings of the 1996 international conference on Learning sciences*, pages 75–82. International Society of the Learning Sciences, 1996. ISBN 1-880094-23-1.

Claudia Gama. Metacognition in interactive learning environments: The reflection assistant model. *Intelligent Tutoring Systems*, pages 668–677, 2004a.

Claudia Amado Gama. *Integrating Metacognition Instruction in Interactive Learning Environments*. Doctor of philosophy, University of Sussex, Brighton BN1 9RH United Kingdom, 2004b.

Mary Gick and Keith Holyoak. The cognitive basis of knowledge transfer. In S.M. Cormier and J.D. Hagman, editors, *Transfer of learning: Contemporary research and applications*, pages 9 – 46. Academic Press, New York, 1987.

Mary Gick and Keith Holyoak. Schema induction and analogical transfer. *Cognitive Psychology*, 15:1 – 38, 1983.

Joesph Gliem and Rosemary Gliem. Calculating, interpreting, and reporting cronbach's alpha reliability coefficient for likert-type scales. *Midwest Research to Practice in Adult, Continuing and Community Education*, 2003.

Bradley Goodman, Amy Soller, Frank Linton, and Robert Gaimari. Encouraging student reflection and articulation using a learning companion. *International Journal of Artificial Intelligence in Education*, 9:237–255, 1998.

R.E. Goska and P.L. Ackerman. An aptitude-treatment interaction approach to transfer within training. *Journal of Educational Psychology*, 88:249 – 259, 1996.

A. Graesser, P. Chipman, B.C. Haynes, and A. Olney. Autotutor: An intelligent tutoring system with mixed-initiative dialogue. *IEEE Transactions in Education*, 48(4):612–618, 2005.

A.C. Graesser, G.T. Jackson, E.C. Mathews, H.H. Mitchell, A. Olney, M. Ventura, P. Chipman, D.Franceschetti, X. Hu, M.M. Louwerse, N.K. Person, and The Tutoring Research Group. Why/autotutor: A test of learning gains from a physics tutor with natural language dialog, 2003.

Arthur Graesser, Danielle McNamara, Max Louwerse, and Zhiqiang Cai. Coh-metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, and Computers*, 36: 193–202, 2004.

Arthur Graesser, Moongee Jeon, Moongee, Yan Yan, and Zhiqiang Cai. Discourse cohesion in text and tutorial dialogue. *Information Design Journal*, 15(3):199–213, 2007.

Arthur C. Graesser, Natalie Person, and Joseph P. Magliano. Collaborative dialogue patterns in naturalistic one-to-one tutoring. *Applied Cognitive Psychology*, 9:495 – 522, 1995.

Barbara Grosz and Candy Sidner. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12:175 – 204, 1986.

Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21:203–226, June 1995.

Richard R. Hake. Design-based research in physics education research: A review. In A.E. Kelly, R.A. Lesh, and J.Y. Baek, editors, *Handbook of Design Research Methods in Mathematics, Science, and Technology Education*. Erlbaum, 2007.

M. A. K. Halliday and Ruqaiya Hasan. *Cohesion in English*. English Language Series. Pearson Education Limited, 1976.

Ibrahim Halloun and David Hestenes. Common sense concepts about motion. *American Journal of Physics*, 53(11), 1985a.

Ibrahim Halloun and David Hestenes. The initial knowledge state of college physics students. *American Journal of Physics*, 53(11):1043–1055, 1985b.

Marti Hearst. Multi-paragraph segmentation of expository text. In *32nd. Annual Meeting of the Association for Computational Linguistics*, pages 9–16, New Mexico State University, Las Cruces, New Mexico, 1994.

David Hestenes, Malcolm Wells, and Gregg Swackhamer. Force concept inventory. *The Physics Teacher*, 30(3):141–158, 1992.

D. Higgins, J. Burstein, D. Marcu, and C. Gentile. Evaluating multiple aspects of coherence in student essays. *Proceedings of the Annual Meeting of HLT/NAACL*, pages 185 – 192, 2004.

G.T. Jackson, N.K. Person, and A.C.Graesser. Adaptive tutorial dialogue in autotutor. In *Proceedings of the workshop on Dialog-based Intelligent Tutoring Systems at the 7th International conference on Intelligent Tutoring Systems*, pages 9–13. Universidade Federal de Alagoas, Brazil, 2004.

Moongee Jeon and Roger Azevedo. Analyzing human tutorial dialogues for cohesion and coherence during hypermedia learning of a complex science topic. In *Proceedings of the 29th Annual Meeting of the Cognitive Science Society*, pages 1127–1132, 2007.

Moongee Jeon and Roger Azevedo. Automatic analyses of cohesion and coherence in human tutorial dialogues during hypermedia: A comparison among mental model jumpers. In *ITS '08: Proceedings of the 9th international conference on Intelligent Tutoring Systems*, pages 690–692, 2008. ISBN 978-3-540-69130-3. doi: http://dx.doi.org/10.1007/978-3-540-69132-7_79.

Pamela Jordan, Brian Hall, Michael Ringenberg, Yui Cui, and Carolyn Rosé. Tools for authoring a dialogue agent that participates in learning studies. In *Proceedings of Artificial Intelligence in Education, AIED*, pages 43–50, 2007.

S. Kalyuga and P. Ayres. The expertise reversal effect. *Educational Psychologist*, 38:23 – 31, 2003.

Sandra Katz, Alan Lesgold, Edward Hughes, Daniel Peters, Gary Eggan, Maria Gordin, and Linda Greenberg. Sherlock 2: An intelligent tutoring system built on the lrdc framework. In C. P. Bloom and R. B. Loftin, editors, *Facilitating the development and use of interactive learning environments*, pages 227 – 258. Erlbaum, Hillsdale, New Jersey, 1998.

Sandra Katz, David Allbritton, and John Connelly. 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, 2003.

Sandra Katz, John Connelly, and Christine Wilson. Out of the lab and into the classroom: An evaluation of reflective dialogue in andes. In *Proceeding of the 2007 conference on Artificial Intelligence in Education*, pages 425–432, Amsterdam, The Netherlands, 2007. IOS Press. ISBN 978-1-58603-764-2.

Cynthia Kersey, Barbara Di Eugenio, Pamela Jordan, and Sandra Katz. Knowledge co-construction and initiative in peer learning interactions. In *AIED 2009, The 14th International Conference on Artificial Intelligence in Education. Brighton, UK*, Brighton, UK, July, 2009.

A.N. Kluger and A. DeNisi. The effects of feedback interventions on performance: Historical review, a meta-analysis and a preliminary feedback intervention theory. *Psychological Bulletin*, 119:254 – 284, 1996.

C. Kulik, J. Kulik, and R. Bangert-Drowns. Effectiveness of mastery learning programs: A meta-analysis. *Review of Educational Research*, 60:265 – 306, 1990.

Eleni A. Kyza, Ravit Golan, Brian Reiser, and Daniel Edelson. Reflective inquiry: enabling group self-regulation in inquiry-based science using the progress portfolio tool. *Proceedings of the Conference on Computer Support for Collaborative Learning: Foundations for a CSCL Community*, pages 227–236, 2002.

Mirella Lapata and Regina Barzilay. Automatic evaluation of text coherence: Models and representations. *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, pages 1085 – 1090, 2005.

Adrienne Lee and Laura Hutchison. Improving learning from examples through reflection. *Journal of Experimental Psychology: Applied*, 4:187–210, 1998.

R. Leher and J. Littlefield. Relationships among cognitive components in logo learning and transfer. *Journal of Educational Psychology*, 85:317 – 330, 1993.

Xiaodong Lin and James Lehman. Supporting learning of variable control in a computer-based biology environment: Effects of prompting college students to reflect on their own thinking. *Journal of Research in Science Teaching*, 36:837–858, Sept. 1999.

D. Litman and S. Silliman. ITSPOKE: An intelligent tutoring spoken dialogue system. In *Companion Proc. of the Human Language Technology Conf: 4th Meeting of the North American Chap. of the Assoc. for Computational Linguistics*, 2004.

Diane Litman and Kate Forbes-Riley. Correlations betweeen dialogue acts and learning in spoken tutoring dialogues. In *Natural Language Engineering*, volume 12, pages 161–176, June 2006a.

Diane Litman and Kate Forbes-Riley. Predicting student emotions in computer-human tutoring dialogues. In *ACL '04: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 351, Morristown, NJ, USA, 2004. Association for Computational Linguistics. doi: http://dx.doi.org/10.3115/1218955.1219000.

Diane J. Litman and Kate Forbes-Riley. Recognizing student emotions and attitudes on the basis of utterances in spoken tutoring dialogues with both human and computer tutors. In *Speech Communication*, volume 48 (5), pages 559–590, May 2006b.

Diane J. Litman, Carolyn P. Rose, Kate Forbes-Riley, Kurt VanLehn, Dumisizwe Bhembe, and Scott Silliman. Spoken versus typed human and computer dialogue tutoring. *International Journal of Artificial Intelligence in Education*, 16:145 – 170, 2006.

Edward Loper and Steven Bird. Nltk: The natural language toolkit. In *In Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics. Philadelphia: Association for Computational Linguistics*, 2002.

W.C. Mann and S.A. Thompson. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8:243 – 281, 1988.

Mark A. McDaniel, Paula J. Waddill, Kraig Finstad, and Tammy Bourg. The effects of text-based interest on attention and recall. *The Journal of Educational Psychology*, 92(3):492–502, 2000.

Kathleen R McKeown. *Text generation : using discourse strategies and focus constraints to generate natural language text*. Cambridge University Press, Cambridge [Cambridgeshire], 1985.

Margaret McKeown, Isabel Beck, Gale Sinatra, and Jane Loxterman. The contribution of prior knowledge and coherent text to comprehension. *Reading Research Quarterly*, 27:79–93, 1992.

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

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

Danielle S. McNamara and Walter Kintsch. Learning from text: Effects of prior knowledge and text coherence. *Discourse Processes*, 22:247–287, 1996.

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. Introduction to wordnet: An on-line lexical database. *International Journal of Lexicography (special issue)*, 3 (4):235–312, 1990.

E. Miltsakaki and K. Kukich. Evaluation of text coherence for electronic essay scoring systems. *Natural Language Engineering*, 10:25 – 55, 2004.

Jane Morris and Graeme Hirst. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17:21–48, 1991.

Barack Obama. Remarks by the president at the national academy of sciences. "http://www.whitehouse.gov/the_press_office/Remarks-by-the-President-at-the-National-Academy-of-Sciences-Annual-Meeting/", 2009.

M. O'donnell, C. Mellish, J. Oberlander, and A. Knott. Ilex: an architecture for a dynamic hypertext generation system. *Natural Language Engineering*, 7(3):225–250, 2001. ISSN 1351-3249. doi: http://dx.doi.org/10.1017/S1351324901002698.

Andrew Olney and Zhiqiang Cai. An orthonormal basis for topic segmentation in tutorial dialog. In *Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*, pages 971–978. Vancouver, October 2005.

Tenaha O'Reilly and Danielle McNamara. Reversing the reverse cohesion effect: good texts can be better for strategic, high-knowledge readers. *Discourse Processes*, 43:121–152, 2007.

Fred Paas, Juhani Tuovinen, Huib Tabbers, and Pascal Van Gerven. Cognitive load measurement as a means to advance cognitive load theory. 38(1):63–71, 2003.

Paul Pintrich and Elisabeth DeGroot. Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology*, 82(1):33–40, 1990.

H. Pon-Barry, B. Clark, K. Schultz, E.O. Bratt, S. Peters, and D. Haley. Contextualizing reflective dialogue in a spoken conversational tutor. *Educational Technology & Society*, 8:42 – 51, 2005.

M. F. Porter. An algorithm for suffix stripping. *Readings in information retrieval*, pages 313–316, 1997.

Stephen Provasnik, Patrick Gonzales, and David Miller. U.s. performance across international assessments of student achievement: Special supplement to the condition of education 2009 (nces 2009-083). *National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.*, 2009.

PSLC. Pittsburgh science of learning center, 2009. URL http://www.learnlab.org/.

R Development Core Team. R: A language and environment for statistical computing, 2005. URL http://www.R-project.org. ISBN 3-900051-07-0.

S.K. Reed. A schema-based theory of transfer. In D. K. Detterman and R. J. Sternberg, editors, *Transfer on trial: Intelligence, cognition, and instruction*, pages 39 – 67. Ablex Publishing Company, Norwood, New Jersey, 1993.

Philip Resnik. Using information content to evaluate semantic similarity in a taxonomy. *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pages 448 – 453, August 1995.

Maxwell Roberts and Riccardo Russo. *A student's guide to Analysis of Variance*. Routledge, 29 West 35th Street, New York, NY 10001, 1999.

Doug Rohde. Linger: a flexible platform for language processing experiments. http://tedlab.mit.edu/ dr/Linger/, 2003.

Ido Roll. *Structured Invention Tasks to Prepare Students for Future Learning: Means, Mechanisms, and Cognitive Processes*. Doctor of philosophy, Carnegie Mellon University, 5000 Forbes Ave. Pittsburgh, Pa., 2009.

Carolyn P. Rosé. A framework for robust semantic interpretation. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, pages 311–318, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc.

Carolyn P. Rosé, Pamela Jordan, Michael Ringenberg, Stephanie Siler, Kurt Vanlehn, and Anders Weinstein. Interactive conceptual tutoring in atlas-andes. In *In*, pages 256–266. Press, 2001.

Susana Rubio, Eva Daz, and Jess Martn. A comparison of swat, nasa-tlx, and workload profile methods. *APPLIED PSYCHOLOGY: AN INTERNATIONAL REVIEW*, 53:61–86, 2004.

Ulrich Schiefele. Topic interest, text representation, and quality of experience. *Contemporary Educational Psycholgy*, 21:3–18, 1996.

Ulrich Schiefele and Andreas Krapp. Topic interest and free recall of expository text. *Learning and Individual Differences*, 8(2):141–160, 1996.

D.A. Schön. *The Reflective Practioner. How Professionals Think in Action*. Templesmith, London, 1983.

Holger Schultheis and Anthony Jameson. Assessing cognitive load in adaptive hypermedia systems: Physiological and behavioral methods. In Wolfgang Nejdl and Paul De Bra, editors, *Adaptive Hypermedia and Adaptive Web-Based Systems: Proceedings of AH 2004*, pages 225–234. Springer, Berlin, 2004.

G.R. Skanes, A.M. Sullivan, E.J. Rowe, and E. Shannon. Intelligence and transfer: Aptitude by treatment interactions. *Journal of Educational Psychology*, 66:563 – 568, 1974.

John Sweller. Cognitive load theory, learning difficulty and instructional design. *Learning and Instruction*, 4:295–312, 1994.

Josephine Tchetagni, Roger Nkambou, and Jacqueline Bourdeau. Explicit reflection in prolog-tutor. *International Journal of Artificial Intelligence in Education (IJAIED)*, 17:169–215, 2007.

Sigmund Tobias. Interest, prior knowledge, and learning. *Review of Educational Research*, 64(1): 37–54, 1994.

T. A. vanDijk and W. Kintsch. *Strategies of Discourse Comprehension*. New York, Academic Press, 1983.

K. VanLehn, A.C. Graesser, G.T. Jackson, P. Jordan, A. Olney, and C.P. Rose. When are tutorial dialogues more effective than reading? *Cognitive Science*, 31:3 – 62, 2007.

Kurt VanLehn, Reva Freedman, Pamela Jordan, Charles Murray, Remus Osan, Michael Ringenberg, Carolyn Rosé, Kay Schulze, Robert Shelby, Donald Treacy, Anders Weinstein, and Mary Wintersgill. Fading and deepening: The next steps for andes and other model-tracing tutors. In *ITS '00: Proceedings of the 5th International Conference on Intelligent Tutoring Systems*, pages 474–483, London, UK, 2000. Springer-Verlag. ISBN 3-540-67655-4.

Kurt VanLehn, Pamela W. Jordan, Carolyn P. Rosé, Dumisizwe Bhembe, Michael Boettner, Andy Gaydos, Maxim Makatchev, Umarani Pappuswamy, Michael Ringenberg, Antonio Roque, Stephanie Siler, and Ramesh Srivastava. The architecture of why2-atlas: A coach for qualitative physics essay writing. In *Proc. 6th Int. Conf. on Intelligent Tutoring Systems*, volume 2363 of *LNCS*, pages 158–167. Springer, 2002. URL citeseer.ist.psu.edu/vanlehn02architecture.html.

Kurt VanLehn, Stephanie Siler, Charles Murray, Takashi Yamauchi, and William .B Baggett. Why do only some events cause learning during human tutoring? *Cognition and Instruction*, 21: 209–249, 2003.

Arthur Ward and Diane Litman. Cohesion and learning in a tutorial spoken dialog system. In *Proceedings of the 19th International FLAIRS Conference (FLAIRS-19)*, pages 533–538, May 2006.

Arthur Ward and Diane Litman. Semantic cohesion and learning. In *Proceedings 9th International Conference on Intelligent Tutoring Systems (ITS)*, pages 459–469, Ann Arbor, June 2008.

Arthur Ward, John Connelly, Sandra Katz, Diane Litman, and Christine Wilson. Cohesion, semantics and learning in reflective dialog. In *Proceedings of the AIED Workshop on Natural Language Processing in Support of Learning: Metrics, Feedback and Connectivity*, June 2009.