THE PATHWAY ACTIVE LEARNING ENVIRONMENT: AN INTERACTIVE WEB-BASED TOOL FOR PHYSICS EDUCATION

by

CHRISTOPHER MATTHEW NAKAMURA

B.S., The University of Michigan, 2003
M.S., Kansas State University, 2006

AN ABSTRACT OF A DISSERTATION

submitted in partial fulfillment of the requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Curriculum and Instruction
College of Education

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2012

Abstract

The work described here represents an effort to design, construct, and test an interactive online multimedia learning environment that can provide physics instruction to students in their homes. The system was designed with one-on-one human tutoring in mind as the mode of instruction. The system uses an original combination of a video-based tutor that incorporates natural language processing video-centered lessons and additional illustrative multimedia. Our Synthetic Interview (SI) tutor provides pre-recorded video answers from expert physics instructors in response to students' typed natural language questions. Our lessons cover Newton's laws and provide a context for the tutoring interaction to occur, connect physics ideas to real-world behavior of mechanical systems, and allow for quantitative testing of physics. Additional multimedia can be used to supplement the SI tutors' explanations and illustrate the physics of interest. The system is targeted at students of algebra-based and concept-based physics at the college and high school level. The system logs queries to the SI tutor, responses to lesson questions and several other interactions with the system, tagging those interactions with a username and timestamp. We have provided several groups of students with access to our system under several different conditions ranging from the controlled conditions of our interview facility to the naturalistic conditions of use at home. In total nearly two-hundred students have accessed the system. To gain insight into the ways students might use the system and understand the utility of its various components we analyzed qualitative interview data collected with 22 algebra-based physics students who worked with our system in our interview facility. We also performed a descriptive analysis of data from the system's log of user interactions. Finally we explored the use of machine learning to explore the possibility of using automated assessment to augment the interactive capabilities of the system as well as to identify productive and unproductive use patterns. This work establishes a proof-of-concept level demonstration of the feasibility of deploying this type of system. The impact of this work and the possibility of future research efforts are discussed in the context of Internet technologies that are changing rapidly.

THE PATHWAY ACTIVE LEARNING ENVIRONMENT: AN INTERACTIVE WEB-BASED TOOL FOR PHYSICS EDUCATION

by

CHRISTOPHER MATTHEW NAKAMURA

B.S., The University of Michigan, 2003
M.S., Kansas State University, 2006

A DISSERTATION

submitted in partial fulfillment of the requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Curriculum and Instruction
College of Education

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2012

Approved by:

Major Professor
Dean A. Zollman

Abstract

The work described here represents an effort to design, construct, and test an interactive online multimedia learning environment that can provide physics instruction to students in their homes. The system was designed with one-on-one human tutoring in mind as the mode of instruction. The system uses an original combination of a video-based tutor that incorporates natural language processing video-centered lessons and additional illustrative multimedia. Our Synthetic Interview (SI) tutor provides pre-recorded video answers from expert physics instructors in response to students' typed natural language questions. Our lessons cover Newton's laws and provide a context for the tutoring interaction to occur, connect physics ideas to real-world behavior of mechanical systems, and allow for quantitative testing of physics. Additional multimedia can be used to supplement the SI tutors' explanations and illustrate the physics of interest. The system is targeted at students of algebra-based and concept-based physics at the college and high school level. The system logs queries to the SI tutor, responses to lesson questions and several other interactions with the system, tagging those interactions with a username and timestamp. We have provided several groups of students with access to our system under several different conditions ranging from the controlled conditions of our interview facility to the naturalistic conditions of use at home. In total nearly two-hundred students have accessed the system. To gain insight into the ways students might use the system and understand the utility of its various components we analyzed qualitative interview data collected with 22 algebra-based physics students who worked with our system in our interview facility. We also performed a descriptive analysis of data from the system's log of user interactions. Finally we explored the use of machine learning to explore the possibility of using automated assessment to augment the interactive capabilities of the system as well as to identify productive and unproductive use patterns. This work establishes a proof-of-concept level demonstration of the feasibility of deploying this type of system. The impact of this work and the possibility of future research efforts are discussed in the context of Internet technologies that are changing rapidly.

Table of Contents

List of Figures........ix

List of Tables........x

Acknowledgements........xi

Chapter 1 - Introduction........1

1.1 Background & Motivation........2

1.2 Research Questions........8

1.3 Research Approach........9

1.4 Broader Impact and Implications........11

1.5 Dissertation Organization........12

Chapter 2 - Review of Relevant Literature........14

2.1 The Constructivist Perspective........14

2.1.1 Piaget and Constructivism........15

2.1.2 Vygotsky and the Zone of Proximal Development........18

2.1.3 The Learning Cycle........19

2.1.4 Ideas from Physics Education Research........22

2.2 Related Work on Human Tutoring........24

2.2.1 The “Two-sigma Problem”........24

2.2.2 Explaining the Effectiveness of Human Tutoring........26

2.3 Related Work on Human-Computer Tutoring........28

2.3.1 Computer-Human Coaching Research........28

2.3.2 A Social Component of Human-Computer Interaction........29

2.4 Multimedia & Digital Video in Physics Education and Beyond........29

2.4.1 Digital Video in Physics Education........29

2.4.2 Prior Research on Multimedia in Instruction........33

Chapter 3 - Research Design and Methodology........36

3.1 Populations of Interest........36

3.2 The Pathway Active Learning Environment........37

3.2.1 The Synthetic Interview Tutor........38

3.2.2 The Pathway Active Learning Environment Lesson Materials 41

3.2.3 Pathway Active Learning Environment Support Multimedia....................................50

3.3 Research Designs and Data Collection Methods 50

3.4 Sampling Methods 51

3.5 Data Analysis Methods...53

3.5.1 Qualitative Analysis of Interview Data 53

3.5.2 Descriptive Analysis of PALE Log Data..54

3.5.3 Using Machine Learning to Group Student Responses 55

3.6 Summary...59

Chapter 4 - Understanding Student Experiences with the Pathway Active Learning Environment 61

4.1 Analysis of Data Logs and Interview Data...61

4.1.1 Student Perceptions of and Interactions with the SI Tutor 62

4.1.2 Analysis of Student Queries to the SI Tutor 66

4.1.3 Students' Perceptions and Use of Video Measurement...69

4.1.4 Quantitative Analysis of Student Use of Video Measurement 71

4.1.4 Student Perceptions and use of Multimedia Support...73

4.2 Discussion of Results 73

4.3 Summary & Conclusions..77

Chapter 5 - Automated Analysis of Short Text Responses...79

5.1 Automated Analysis of the Short Responses to Lesson Questions..................................79

5.1.2 Automated Assessment Analysis Procedure...79

5.1.2 Success Criteria for the Analysis Scheme 83

5.1.3 Results of Automated Analysis Procedure..84

5.3 Discussion of Short Response Data and Analysis 92

5.4 Summary & Conclusions..95

Chapter 6 - Conclusions 96

6.1 Overview of the work..96

6.2 Research Questions Answered...96

6.3 Future Work...98

6.3.1 Modifications to PALE and further work with this system 98

6.3.2 Further Work 99
6.3.3 Understanding this Work in the Context of Emerging Internet Technology............100
6.4 Final Conclusions..102
References 106
Appendix A - PALE Lesson Materials 113

List of Figures

Figure 1.1 The Physics Teaching Web Advisory Interface 3
Figure 1.2 An early design sketch for the synthetic tutoring system..8
Figure 2.1 The three stage learning cycle 20
Figure 3.1 The Pathway Active Learning Environment interface..38
Figure 3.2 Schematic of SI-Tutor's functionality 39
Figure 3.3 Tutors recruited for the project..40
Figure 3.4 Proposed machine learning analysis scheme for student responses to lesson questions 58

List of Tables

Table 2.1 Piaget's Logical Knowledge Stages of Cognitive Development 15

Table 3.1 Lesson one exploration activities 44

Table 3.2 Lesson one application activities.. 45

Table 3.3 Lesson two exploration activities 46

Table 3.4 Lesson two application activities 47

Table 3.5 Lesson three exploration activities 48

Table 3.6 Lesson three application activities 49

Table 3.7 Student sampling for PALE studies 52

Table 4.1 Querying behavior across all three lessons for all interviewed students who had access to the SI Tutor (N=19). 68

Table 4.2 Average number of queries submitted to the SI tutor for the four treatment groups across all three lessons and in total. The indicated error bar is the standard error. 68

Table 5.1 Benchmarking demarcations for Cohen's Kappa values.. 84

Table 5.2 Match rate for machine learning analysis method 85

Table 5.3 Match/mismatch details for lesson 1 exploration activity 1 question 5 88

Table 5.4 Match/mismatch details for lesson 1 exploration activity 3 question 5 89

Table 5.5 Match/Mismatch details for lesson 1 application activity 1 question 1 90

Table 5.6 Match/Mismatch details for lesson 2 application activity 1 question 1 91

Table 5.7 Match/Mismatch details for lesson 3 exploration activity 2 question 3 92

Acknowledgements

I would like to acknowledge the help and support of many without whom I would not have been able to complete this work. Obviously many thanks are due to my advisor, Dean Zollman, who has an amazing ability to strike the right balance between giving advice and encouraging independence. I have learned so much from working with him. Dean, thank you for taking a risk and taking me on as your student. Sanjay Rebello is a remarkable individual who teaches so many of us in the K-SUPER group by advising and by example. I've learned a great deal from talking with him and from watching him work as a researcher and teacher. I feel very fortunate to have worked with amazing people on the Pathway project: Sytil Murphy, Josh Gross, Bryan Maher, Mike Christel, Scott Stevens, Brian Adrian, and Nasser Juma. I also feel fortunate to have gotten to work with so many great people in the K-SUPER group over the years; it has been such a great environment to work in. Two stand out, however as needing special mention: Adrian Madsen who always checks to make sure we're all still alive and Liz Gire who, in such a brief period of time, served as an unexpected role model. I have to thank Kim Coy for all the important things she does to help us conduct our research and drag our posters and papers all across the country. I also have to thank Peggy Matthews for fixing my tuition approximately forty-seven times and generally making life go more smoothly. For all of us, the significance of our lives can be defined only in terms of connections to the lives of others. I have been very fortunate to have benefited from so many helpful connections with friends, former co-workers (almost all of whom are also friends) and family. A huge thanks to my parents for instilling in me the love of learning that started all this trouble, for their love and for generously supporting me in my endeavors in many ways. Thanks to my brother for always being there to talk about interesting things and provide another view on whatever I'm thinking about. Finally, I am forever grateful to Sarah Nuss-Warren for her love and support. This work was supported in part by grants from the United States National Science Foundation under grant numbers REC-0632587 and REC-0632657.

Chapter 1 - Introduction

Reliable research findings on the instructional efficacies of emerging technologies have often lagged behind the claims about the transformative power of these technologies in education. A contemporary example can be seen in discourse surrounding the Khan Academy, an online instructional website based around explanatory videos narrated by founder, Salman Khan (www.khanacademy.org, 2011). Proponents of online instruction have claimed that this system, and others like it will supplant traditional classroom instruction. Steven Pearlstein, for example harshly criticizes the current state of education and concludes, without research-based evidence, that video instruction, like that offered by the Khan Academy, represents a good solution (Pearlstein, 2011). The objective validity of this assertion is unknown, but the author's perceived authority may lend the argument some weight, regardless of its actual validity. The economic implications for parents, teachers and society at large associated with accepting or rejecting these claims are clear and underscore the importance of solid research findings being the underpinnings of decisions about curriculum, and educational policy. Unfortunately Pearlstein does not cite research findings that suggest short narrated videos are superior to classroom instruction (2011). As Internet technologies increasingly become a part of 21st century life we can expect people to increasingly call these technologies into service for education. In order to make good decisions about implementation, we need solid research results that help identify best practices for using these technologies for instruction.

The idea that educators can use video, multimedia and the Internet to positive effect is intuitive, and prior research results that support this idea are presented in this dissertation. These technologies can be used in many different ways and the most interactive and novel of these are only now emerging as the Internet continues to develop. It is not clear how to most effectively combine and use the various available technologies for instruction. Presently, the Internet offers a wealth of interactive capabilities that can be combined with multimedia modes of presentation to build very sophisticated learning environments, but research establishing clear design principles and implementation strategies is lacking. Without these principles and strategies teachers, researchers and other curriculum developers are left with trial and error as the primary means of developing effective web-based instructional materials.

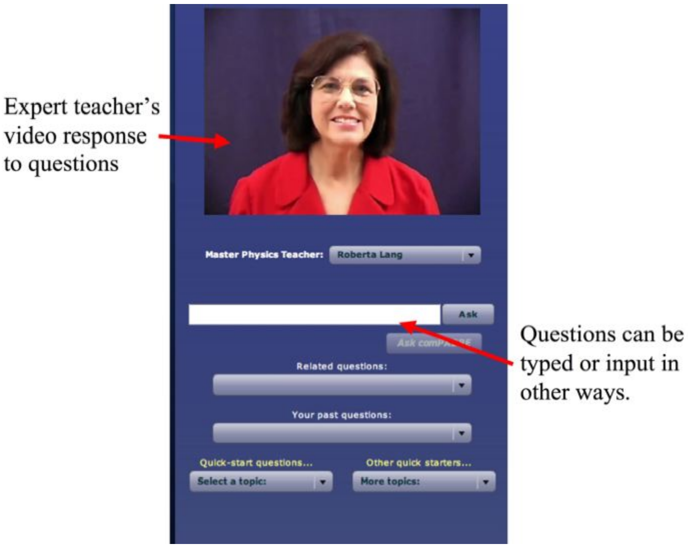
In this dissertation we present research designed to help fill in this gap. At its most general, the fundamental research question addressed by our work is this: what would a modern interactive multimedia-driven learning environment look like? More specifically we have explored this question in the context of physics, which is an active area of discipline-specific education research and provides a natural context for exploring these issues. To address this broad overarching question we tested several combinations of interactive web-based multimedia technologies in an effort to explore the ways that these technologies could be used to positive effect in supporting student learning of Newton's laws by algebra-based and concept-based physics undergraduate physics students as well as high school physics students . At the same time we have explored the assertion that this type of technology can be used to study student learning of physics in those same populations. To do this we have exploited the Internet's ability to quickly deliver multimedia, as well as its interactivity to produce a learning environment that is designed to emulate one-on-one tutoring. Learning frequently a social activity. One-on-one tutoring is both one of the most socially interactive methods of learning, and one of the most effective. Recognizing the benefits of one-on-one tutoring leads us naturally towards incorporating that mode of instruction into our system design. It then becomes part of our goal to explore to what extent it is possible to create an online instructional system that emulates the positive social components of tutoring.

The remainder of this chapter provides an introduction to the research we are presenting, beginning with a previous research project from which this research has evolved. In this chapter we also discuss the motivations for performing this research, the research questions we have investigated, and our general research approach. The anticipated broader impact of our research within the physics education research community and beyond is also discussed. The chapter concludes with a discussion of the organization of the remainder of this dissertation.

1.1 Background & Motivation

This project is a natural evolution of the Physics Teaching Web Advisory (Physics Pathway) project. In that work educator-researchers sought to provide pre-service and in-service high school physics teachers with an electronic resource to help them better teach physics (Stevens, Zollman, Christel & Adrian, 2007). A significant fraction of high school physics teachers are not primarily trained in physics, but must teach physics as part of their appointments

(Neuschatz & McFarling, 2000). At the same time, even beginning teachers whose content knowledge is quite solid, may have little practical experience in how to best teach physics topics. People in these situations may have questions about how to most effectively teach physics, and it would be helpful for them to be able to discuss these issues with a more experienced physics teacher. The Physics Pathway is a website developed to meet this need. With the aid of Synthetic Interview (SI) technology, which is discussed in detail in Chapter 3, section 3.1, novice physics teachers can ask natural language questions regarding physics demonstrations and pedagogy via the keyboard. They then receive pre-recorded video responses to those questions, delivered by expert physics teachers (Stevens et al., 2007). The goal of the SI technology is to lend a socially interactive element to obtaining information on the web. By trading questions and answers the system begins to simulate some of the basic features of a simple conversation (Marinelli & Stevens, 1998). To create an effective help system several master teachers have recorded responses to common questions, and in total, answers to approximately 7,600 distinct questions can be viewed. This SI technology is the central feature of the Physics Pathway. The portion of the system interface that contains and controls the SI is shown in Figure 1.1.


Expert teacher's video response to questions

Questions can be typed or input in other ways.

Figure 1.1 The Physics Teaching Web Advisory Interface

The explanations that teachers receive from the SI interface are then supplemented by additional multimedia in two ways. First, in some instances the master physics teachers felt that their answer could be clarified by an image, and they produced the image, typically by hand.

Because of the difficulty involved in displaying their images within the SI video response, the pictures were digitized so that they could be displayed alongside the expert teacher's response at the appropriate time. In addition to these images, the Physics Pathway provides access to a powerful digital video library that allows teachers to access video clips that reference physics concepts relevant to their questions (Christel, Kanade, Mauldin, Redy, Sribu & Stevens, 1995). The combination of these multimedia components creates a multimodal learning aid that can address multiple learning styles. The system also provides teachers with video clips that they can directly use in their classrooms.

A natural extension of this project is to ask whether a similar online interface could help answer students' questions about physics content. It is not difficult to envision a student having difficulties with homework late at night or on the weekend. At that point the teacher is unavailable. Without a resource, the student's progress may be impeded for the evening. With a resource, the student may be able to get past the difficulty and continue to make progress. Even if the resource does not enable the student to completely resolve the difficulties, at the very least it may both prolong the time spent thinking about the material and change the manner in which the student thinks about the material. In the long run that additional effort and the experience of working with the resource could still help the student improve his or her understanding, particularly when coupled with further instruction. In the past the only resource one could really envision in this type of situation was a textbook. More recently one could envision a search engine on the Internet providing the student with information, but this is still not particularly interactive. It also depends entirely on the student for creating a sense of focus or coherence; the student could easily spend a great deal of time with the search engine while doing very little to improve his or her understanding of the relevant material. The goal of this project is to construct and evaluate a resource that is interactive and can also be of direct help to a student when his or her teacher is unavailable. Our goal is to build and evaluate an online tutoring system based on the SI technology described above and similar in nature to the Physics Pathway that can provide students with video responses to their natural language questions. Those SI video responses can are supplemented with additional images or demonstrative video clips that address the relevant concepts, help the tutor connect physics ideas to physical reality and may help the students better build their physics knowledge. We call this system the Pathway Active Learning Environment (PALE).

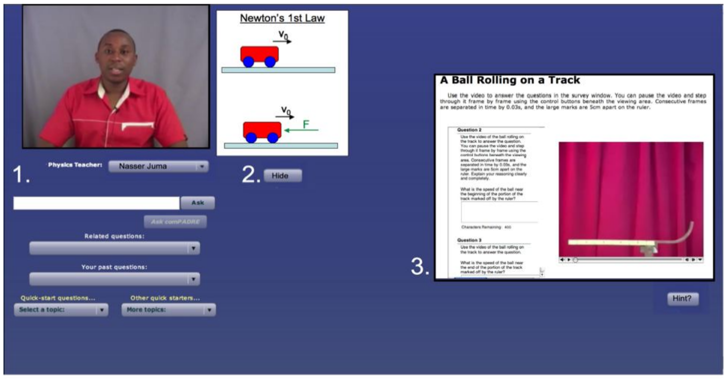
Viewing our system in the context of tutoring provides us with a wealth of tutoring literature to build upon throughout the development process. Tutoring may, in some sense, be the oldest form of instruction. It also may be the most effective (Bloom, 1984). Learning gains similar to tutoring can be achieved by other means, but with significant effort (Bloom, 1984). The ability to achieve the efficacy of tutoring via direct replication of the tutoring intervention would therefore be of significant value to educators.

The social engagement inherent to tutoring may, in part, explain its efficacy as an instructional method (Chi, Siler, Jeong, Yamauchi & Hausmann, 2001; Chi, Siler & Jeong 2004; Okita, Bailenson & Schwartz 2007). One could argue that since students only interact with a video interface the described system is not socially interactive. Prior research has addressed this issue, suggesting simulated social interactions via computer also provide learning benefits (Okita et al., 2007). We discuss that work in detail in Chapter 2. The key finding in Okita et al.'s work is that a student's perception of social interaction may be more important than the presence of a real social interaction (2007). We believe that students, aware that the SI represents a real human who recorded the video, may assign some social weight to the interactions. This idea, that people interact with computers as if they were people is sometimes referred to as the computers as social actors hypothesis (Reeves & Nass, 1986). This hypothesis is one that our system can investigate. One could also argue that since the system we propose requires students to type questions and answers, rather than speaking them as one would do in a normal tutoring session, the system is very different from real tutoring and is lacking in social interaction. However, recent research has shown that students reap similar educational benefits from a computerized tutoring system regardless of whether they type or speak (D'Mello, Dowell & Graesser, 2011). At a more intuitive level we know that online chatting, which is an interaction that requires typing, is quite common and I doubt many would argue either that this is not socially interactive or that it is socially interactive in a very different way than talking face-to-face. Students may perceive elements of social interaction while working with our system, and those elements, should they exist, would likely have commonalities and differences with the social elements of a real tutoring session. One of our research motivations is then to empirically investigate to what extent we can simulate the social interaction of a tutoring session with our SI interface, and how the various multimedia components of our SI tutoring system can affect the system's overall utility.

A key realization in the development of the system is that tutoring does not typically consist only of question-answer exchanges that occur in a vacuum. It is typically conducted within the context of some type of homework or other assignment. While ideally a student could bring his or her own homework to the interaction, clearly this is an overly ambitious first step. Predicting the type of work that the student will bring, the range of content knowledge that will be required to complete the work, and a variety of other information that is necessary to help the student with the work is impossible. For the computer to gather and interpret this information on a case-by-case basis is also essentially impossible. In order to be successful in building a proof-of-principle system and to be successful in encouraging a synthetic tutoring interaction, we will need to begin by providing the lesson materials that provide the context for the interaction. Once we have achieved success at the level, we may begin to consider how to address the case in which the student brings an unknown assignment to the tutoring session.

The lesson materials that we design obviously should provide students with questions, tasks, and problems that build and test their understanding of the relevant content material. At the same time, they should be designed to promote use of the SI-tutor. The lessons which we designed to meet this need will be discussed in detail in Chapter 3; here it is sufficient to note that our system really must consist of three components: the lessons, the SI-based tutor, and supplemental multimedia. An early conceptualization of the system's interface, which shows these three components, is shown in Figure 1.2. Our goal then becomes to investigate the way in which these three components can be brought together to support learning.

Questions about the efficacy of different multimedia and pedagogical constructions flow naturally from this research goal. Research and curriculum development activities focused on the utility of digital video and video measurement and analysis in physics instruction has already established digital video as a useful instructional tool in physics (Brown & Cox, 2009; Laws & Pfister, 1998; Escalada, & Zollman 1997; Zollman & Fuller, 1994). Criticism has been leveled that in the physics education community controlled experiments have not produced sufficient quantitative data to test the efficacy of multimedia in instruction (Lewis, 1995). However, research aimed at experimentally testing and comparing the efficacies of various multimedia modes of presentation in the context of instruction can suffer from a number of severe confounding factors that are not easily overcome. These difficulties include problems controlling extraneous variables within multimedia materials, and subject reactivity (well-known examples include the Hawthorne, John Henry and novelty effects). In essence a conflict exists between constructing an experimental design that is simple enough to execute and simple enough that one can reasonably expect to obtain results that are easily interpretable, and constructing a design that actually reflects the complexities of using multimedia as an instructional tool. These difficulties call into question the utility of simple "this versus that" experimental designs. Correspondingly, a lack of consensus exists on the relative efficacy of different modes of multimedia presentation (Clark, 1983,1994). Clark has gone so far as to suggest that all modes of presentation are fundamentally equivalent (1994). While different modes of presentation can be made equivalent through methodical design, all implementations are clearly not equivalent. We therefore argue that, while simple comparative, experimental designs may not provide great insight into multimedia design, research on multimedia in instruction is- or can be a productive area of study. We further argue that research in this area should focus less on arguments about intrinsic efficacy and design principles and more on naturalistic observation and the development of effective implementation techniques. The research we have conducted centers on methods of effectively combining various multimedia components, including the interactive SI tutor, in ways that promote learning. Therefore, while comparing multimedia is inherent to the research, we do not adopt simple comparative experimental designs. Preliminary work on this project has shown that observing significant differences on a standardized pre-test and post-test is unlikely, but much can be learned from studying multimedia in instruction (Nakamura, Murphy, Zollman, Christel & Stevens, 2010). The research we discuss here does not seek to answer the questions about the efficacy of various forms of multimedia in instruction, which have been unanswered for decades, but instead seeks to establish naturalistic, observational research methods that can serve ongoing programs of study that we hope will ultimately produce richer, more nuanced answers to these lingering questions.



A Ball Rolling on a Track

Use the video to answer the questions in the survey window. You can pause the video and step through it frame by frame using the control buttons beneath the viewing area. Consecutive frames are separated in time by 0.03s, and the large marks are 5cm apart on the ruler.

Question 2

Use the video of the ball rolling on the track to answer the question. You can pause the video and step through it frame by frame using the control buttons beneath the viewing area. Consecutive frames are separated in time by 0.03s, and the large marks are 5cm apart on the ruler. Explain your reasoning clearly and completely.

What is the speed of the ball near the beginning of the portion of the track marked off by the ruler?

Characters Remaining: 400

Question 3

Use the video of the ball rolling on the track to answer the question.

What is the speed of the ball near the end of the portion of the track marked off by the ruler?

1. SI Tutor answers students' physics content questions
2. Supporting multimedia is displayed along side SI Tutor.
3. Lesson materials are displayed on the right.

Figure 1.2 An early design sketch for the synthetic tutoring system

1.2 Research Questions

Beyond the broad overarching question of whether or not we can tutor effectively with an online system, the proposed research is appropriate for addressing several basic research questions. The project is largely exploratory in nature, rather than explanatory. The research questions that we explored with this system are:

- How might students productively, or unproductively, interact with our online tutoring environment?
- How do students feel about working with this type of system? Is this something they perceive to be beneficial?
- How does the introduction and variation of multimedia components within the tutoring environment impact student learning?
- What technological or pedagogical features must be developed and implemented to build a video-based tutoring system?

Other projects that have sought to explore computerized tutoring from different perspectives, such as Andes Tutor and Autotutor, have been ongoing for many years (VanLehn, Lynch, Schulze, Shapiro, Shelby, Taylor & Treacy, 2005; Graesser, Jeon, & Duffy, 2008). Thus, the research proposed in this prospectus is not an attempt to thoroughly document and refine every facet of the system's educational and research merits, but instead an initial effort to develop this type of system and explore some of the most fundamental and accessible features and capabilities, as a basis for further research.

1.3 Research Approach

Because of the exploratory nature of the research questions and the difficulties associated with objective measurement of student understanding and the process of knowledge construction, a mix of qualitative and quantitative research is appropriate. During the project we monitored students' learning as they worked with the system. We looked for ways to establish means of comparing the work of students who have experienced different multimedia instruction while also collecting and examining students' typed responses to free-response questions as a means of developing a richer picture of their understanding of physics.

There are several research contexts or environments appropriate for studying this type of tutoring system. Ultimately, we hope this type of system will be useful to students in their homes. If we are to learn anything from the system when students are using it in that context, we must design it to log as much of the students' interactions with the system as possible. The system was designed to log students' queries to the SI tutor, responses to lesson questions, and user ratings of SI responses. At the same time we cannot know everything about student use of the system in that context. We therefore observed student use of the system in controlled, though artificial, environments. The two obvious choices then become the classroom setting which allows instructors or facilitators to supervise and observe student use and the clinical interview setting, which allows for still more control and observation. As we will discuss in Chapter 3, our data collection focused on these three contexts:

1. Use in a clinical interview setting
2. Classroom use
3. At home use (or use in a location of the student's choosing)

Looking at data from all three settings provided us with the best chance of piecing together an understanding of how our system can be effective. Data consisted of student interactions that can be logged by the system, transcripts from interviews, and video observational data obtained from student interview participants.

Data analysis will be discussed in greater detail in chapter 3, section 3.5. Here we note that the data analysis is both qualitative and quantitative in nature, especially in the beginning of the data analysis process. The data that we have collected are comprised of the log of student interactions with our system (questions to the tutor, responses to lesson questions, etc…), as well as transcripts of interviews conducted with students who used the system in the clinical setting. These types of data can be analyzed from a phenomenographical perspective (Marton, 1981, 1988). The phenomenographical perspective seeks to characterize the different ways in which people experience a given phenomena (Marton, 1981). This perspective on qualitative research is applicable to education because of its intrinsically experiential nature, and its use in that capacity has been explored (Marton, 1981, 1988). Our approach is not a pure phenomenographical design in the sense that we did not attempt to build a hierarchical outcome space, but rather, our approach was informed by the phenomenographical perspective. In our approach students' statements in interviews can be coded line-by-line and analyzed for emergent themes and important instances. These can be brought together to form a coherent picture of students' experiences with the system. By generating multiple lessons and letting students work with the system over the course of multiple weeks we may observe changes in their attitudes or understanding over time.

One approach to data analysis that is more quantitative and that might facilitate the generation of this type of time-resolved picture of students' system usage and learning uses techniques from data mining and natural language processing. In this approach we begin by looking for similarities in the ideas different students express in their responses to a given question. Our observation of these types of groups, may allow for the identification of correlations in student responses across questions, that is identify patterns of responses that characterize differences in how students understand the material. At the same time, we show that it is possible to train computer models to tag responses based these groups. There are many positive implications of a trained computer model being able to effectively code student responses, the most important being that the computerized classification of student responses may facilitate feedback to students, which in turn would help the system function more like a computer. Computer programs appropriate for this type of analysis have already been written and we make use of this prior work (Rosé, Wang, Cui, Arguello, Stegmann, Weinberger & Fischer, 2007). Furthermore, this type of approach has been used in biology education to code essays automatically (Nehm & Haertiz, 2011). It is also beginning to emerge as an interesting means of analyzing short-answer questions in physics education (Butcher & Jordan, 2010;

Nakamura, Murphy, Zollman, Christel & Stevens, 2011; Jordan, 2012) This approach to analysis is discussed in greater detail in Chapter 3, section 3.5 and the results of the analysis are discussed in Chapter 5.

Because of the exploratory nature of this research, the goal of this work is not to produce a theory, or test existing theory, but instead to explore the themes associated with students' interactions with online interactive instructional materials.

Once a qualitative analysis is complete we may have better insight into ways that quantitative techniques can be used to extract more information about student usage patterns from the system's logs. This possibility is also discussed in Chapter 3, section 3.5.

The research is grounded in a constructivist framework. The central ideas is that students must create their own knowledge structures based on their educational experience and that their prior knowledge in a certain sense acts as a seed for that knowledge construction process, and in every sense informs and affects the knowledge construction process. This theoretical grounding is discussed in greater detail in Chapter 2, section 2.1.

1.4 Broader Impact and Implications

The first and most direct opportunity for this proposed research to impact the broader communities of physics education research and physics instruction is in the sense that the lessons we learn in developing this proof-of-principle implementation may be applied to the generation of a more advanced learning environment that is useful for helping students to better learn across an entire curriculum. The Physics Pathway project has now expanded to address teachers' questions on all of mechanics, electrodynamics, and much of modern physics. If this work shows that our system is useful for teaching students about kinematics and Newton's laws, then it may be productive to expand the lessons to cover much of the same material addressed by the Physics Pathway system.

Beyond this purely pedagogical impact the expectation is always that research will produce results that inform further research. There are several ways in which this project has produced result that are of interest for future research projects. This research provides insight and guidance for the best methods of developing this type of technology for student learning, and for implementing this type of technology in ways that promote student learning, and in ways that avoid potential pitfalls of the misuse of technology.

1.5 Dissertation Organization

The remainder of this dissertation will discuss the research that has been done, in greater detail. Chapter 2 will discuss prior research relevant for this study, including theoretical foundations, most notably the constructivist theories of learning, as well as review of work that has set these theories in a context more suitable for physics education research. Relevant literature on useful knowledge organization schemes and cognition models will also be discussed in Chapter 2. Empirical and theoretical research that has bearing on this work will also be discussed. This research mostly focuses on prior studies of tutoring, both purely human tutoring, and human-computer tutoring. Assessment in physics education research will also be discussed. Chapter 2 will close with discussions of research focused on the instructional efficacy of multimedia, including multimedia design theory and cognitive load theory which are related to this proposed work, not as much as a theoretical groundwork but as a research domain on which this work might have potential positive impact.

Chapter 3 discusses the methodology that will be used to conduct the research, including an in-depth discussion of the technology developed for this proposed research, the populations of interest, research designs, the methods of obtaining samples, data collection methods, as well as data analysis techniques.

In Chapter 4 the results of a quantitative analysis of data extracted from the PALE data logs as well as interview data collected with general physics students will be presented. This analysis is qualitative in nature and focuses on students' experiences as they worked with the system.

Chapter 5 presents the results of an analysis scheme for typed responses to the short-answer questions collected through the PALE lesson activities. This approach combines human classification of text responses with analysis via computer models trained on human classification schemes to investigate the possibility of automated analysis of future responses. This approach may enable us to provide more detailed, specific feedback to students based on their actual use of the PALE system. The approach may be of interest to designers of online learning environments more generally because of its potential for enabling feedback without constant human intervention.

Chapter 6 concludes the dissertation and presents a summative discussion of the research questions answered, the potential for future research, and final conclusions that can be drawn from this research effort.

Chapter 2 - Review of Relevant Literature

The literature relevant to our project can be divided into three components. The first component is literature that forms a theoretical groundwork upon which our work is based. Most prominent amongst this material are the constructivist theories of Piaget and Vygotsky. The second major component comprised of research specifically related to tutoring, and human-computer interaction. These studies allow our present research to connect to a web of related research efforts, and ultimately, to inform the future generation of theory and empirical studies in the field of computer-based tutoring. The third component is comprised of previous work on digital video usage in physics education as well as work on multimedia in instruction more generally. The former will focus on applications to give the reader a sense of what has already been done. The latter will center on a discussion of cognitive load theory and multimedia design theory, which is relevant to research in that researchers in those fields have established ideas on what constitutes effective use of multimedia in instruction but, for reasons that will be discussed in detail, we do not use this work as a theoretical grounding for ours.

2.1 The Constructivist Perspective

Efforts to understand learning in individuals have gradually shifted from being largely grounded in philosophy to being more scientifically grounded in the discipline of psychology (Vygotsky, 1997). Similarly psychology has increasingly shifted from a mind-centered orientation to a brain-centered orientation (Vygotsky, 1997). This shift reflects a scientific understanding of humans as wholly biological entities whose mental and physical existences are intimately intertwined. Clear evidence for these transitions can be seen during the first half of the twentieth century. An important result of scientific efforts to understand learning is the constructivist perspective on learning.

The constructivist perspective on learning has, at its core, the fundamental assumption that the process of learning is the process of constructing and organizing knowledge based on life experience. Directly following this assumption in importance is the assertion that our prior knowledge exerts great influence on how we construct and organize new knowledge. Two of the main figures whose independent work informs constructivist theories of learning are Jean Piaget and Lev Vygotsky. Their theories differ somewhat in focus, but here we adopt the perspective that these differences are not contradictory, but instead reflect different facets of the knowledge construction process. Their respective views on knowledge construction will be discussed in turn, beginning with Piagetian constructivism.

2.1.1 Piaget and Constructivism

Piaget's work as an epistemologist, studying learning processes in children and adolescents is most associated with constructivist theories of learning that emerged from his work (Beilin, 1992). Piaget's perspectives on learning evolved through multiple stages (Beilein, 1992). We, however do not adopt the entire body of his work as grounding for our research. We are primarily interested in the facets of Piagetian constructivism that deal with formal reasoning and application of the ideas of adaptation and assimilation to the modification of knowledge structure (Flavell, 1996). Researching children's intellectual development, it was natural for developmental staging to be a naturally emergent idea in his work (Beilin, 1992; Inhelder & Piaget, 1958). Useful logical knowledge stages of cognitive development identified by Piaget and the different characteristics of thinking that typically occur within those stages are described in Table 2.1 (Fuller, Campbell, Dykstra & Stevens 2009).

Table 2.1 Piaget's Logical Knowledge Stages of Cognitive Development

Stage	Age (Years)	Characteristics
Sensorimotor	0 to 1	Pre-verbal reasoning
Pre-operational	1 to 8	No cause and effect reasoning. Uses verbal symbols, simple classifications, lacks conservation reasoning
Concrete Operational	8 to ?	Reasoning is logical but concrete rather than abstract
Formal Operational	11 to ?	Hypothetico-deductive reasoning

Piaget characterizes these stages of learning based on the level of abstract of thinking that a child is capable of during that stage of development (Inhelder & Piaget, 1958; Fuller et al., 2009). Piaget identified these stages by observing children working through different tasks that required certain levels of reasoning ability, and determining the ages at which those reasoning abilities began to appear (Inhelder & Piaget, 1958). Researchers in physics education have observed that the concrete and formal operational classifications to be useful ideas even in adult populations (McKinnon & Renner, 1971). McKinnon and Renner found that about 50% of 131 college freshmen exhibited concrete operational behavior patterns when given tasks developed by Inhelder and Piaget to evaluate formal reasoning (1971). Only 25% of the students in that study operated at the formal operational level. The remainder exhibited a "Post-Concrete" level of thinking which is somewhere in between Concrete and Formal operational states. This finding begs the questions of whether students can become more formal in their thinking, and whether the study of subjects such as physics will promote such growth. McKinnon and Renner did observe changes in students' logical thinking over time. In the same study, they contrasted the formal reasoning of a group that received an inquiry-based treatment curriculum with that of a control group, which had not received instruction. They found gains in the number of students who functioned at the formal operational level in both groups. However, they observed significantly larger gains for the group that was taught with the inquiry-based curriculum. Piaget and colleagues, however never worked with adult populations, and may have been unaware of the possibility that adults could exhibit traits of concrete operational reasoning. McKinnon and Renner's work suggests that Piaget's Concrete and Formal Operational categories are useful for describing student reasoning in adult populations. Even within pre-adult populations there are questions about the universality of the scheme. Hundeide conducted research on student reasoning that indicated that the level at which children operated depended critically on the methods of investigation (1977). We suggest that though the age ranges outlined by Piaget and colleagues may provide a useful framework for understanding reasoning in some children, they can be rejected when dealing with adults without rejecting other relevant facets of their work. In particular, we suggest that the levels of reasoning may be meaningfully applied without the age ranges.

An important concept in Piagetian constructivism is the idea of a schema (Inhelder & Piaget, 1958). We can think of a schema as a conceptualization of a person's knowledge as well as its structure and organization. Schemata can be modified either by assimilation or accommodation. In the case of the former, new information, which is consistent with prior knowledge is added into a person's schema. In the case of the later the schema is more radically altered to accommodate information that is in some way inconsistent with prior knowledge. In general assimilation is much easier than accommodation (Reddish, 2003). Accommodation generally requires a form of motivation on the part of the student (Reddish, 2003). The student must recognize a problem with the current understanding and a compelling reason to change to a better understanding (Reddish, 2003). It is the two-component assimilation-accommodation model of learning more than any other idea that places emphasis on an active, constructivist perspective on learning, and this may be Piaget's most lasting contribution to educational psychology (Flavell, 1992).

This process is related to the idea of cognitive dissonance, a feeling of discomfort due to a schism either between a person's knowledge and an external observation (which is typically undeniable in nature) or a schism between two or more beliefs of which the person has previously been unaware (Festinger, 1957). Obviously, the cognitive dissonance can be addressed in two ways. One can avoid ideas that are in conflict with one's own beliefs, or one can change one's beliefs to conform to external observations. Much of the research on cognitive dissonance has focused on self-image and the rationalization of actions that may be in conflict with a person's self-image; a good example being the action of smoking in spite of the facts that it shortens life, and most people want to live a healthy and long life (Freeman, Hennessy, & Marzullo 2001). This view is not inconsistent with an application of the concept in education. For example, the clear demonstration of a conflict between how a student describes an object's motion and how Newtonian dynamics describes an object's motion can produce two internal conflicts. Clearly, the conflict is between the student's understanding of the situation and science's accepted description of the situation. However, the conflict is also between the student's self-image as a knowledgeable person who understands how things work, and a demonstrated image of someone who does not view the world in a way that is consistent with Newtonian dynamics, a well-accepted model of physical behavior.

Two methods of addressing cognitive dissonance were mentioned: changing beliefs and avoiding the source of conflict. It is therefore in some sense logical, though from our standpoint undesirable, to conclude that if the study of physics produces this conflict, one should stop studying physics. We therefore argue that teachers, instructional materials and other pedagogical interventions must carefully strike a balance in which students' conceptions are elicited and confronted, but in ways that do not drive students away from the study of physics.

One technique that has been devised to confront beliefs that are in conflict with accepted theory in a relatively innocuous manner is the Predict-Observe-Explain task that has been promoted by White and Gunstone (1992). The technique was first used at the University of Pittsburgh as Describe-Observe-Explain (Champagne, Klopfer & Anderson, 1980). In either case, the key to this technique is to elicit student predictions and compare, contrast those predictions with clearly observable phenomena and generate an explanation of the phenomenon, which is in concert with accepted theory. This work has bearing on our project directly, in the sense that we should use this type of activity in our lesson materials to encourage students to build knowledge that is consistent with the physical world. It will be asserted, with deeper discussion, in section 2.2.2 that this type of task is also useful for simulating some of the actions of tutors.

Intuitively, the process of knowledge construction can only occur if the student has means of interaction with the outside environment. Much of the interaction associated with learning consists of social interactions between people. In the next section we address how Vygotsky's work on social learning connects to the constructivist foundation laid in this section.

2.1.2 Vygotsky and the Zone of Proximal Development

Like Piaget, Lev Vygotsky's work focused on developmental psychology and learning in individuals, but with a greater emphasis on social aspects of learning than knowledge structure and organization (1978). A key concept introduced by Vygotsky to explain learning as a social process is the Zone of Proximal Development (ZPD) (Vygotsky, 1978). The ZPD is defined as the range of achievement, which is made accessible to a student by aid from a more knowledgeable person (Vygotsky, 1978 p. 86). Therefore the ZPD naturally excludes the students' current abilities, since those do not require a more knowledgeable person to become apparent. It also excludes a range of achievement that the student cannot attain even with more knowledgeable aid. With this concept in place the process of learning is mediated through social interactions which move achievements within the ZPD into the range of achievement which the student can accomplish alone, and achievements beyond the ZPD into it.

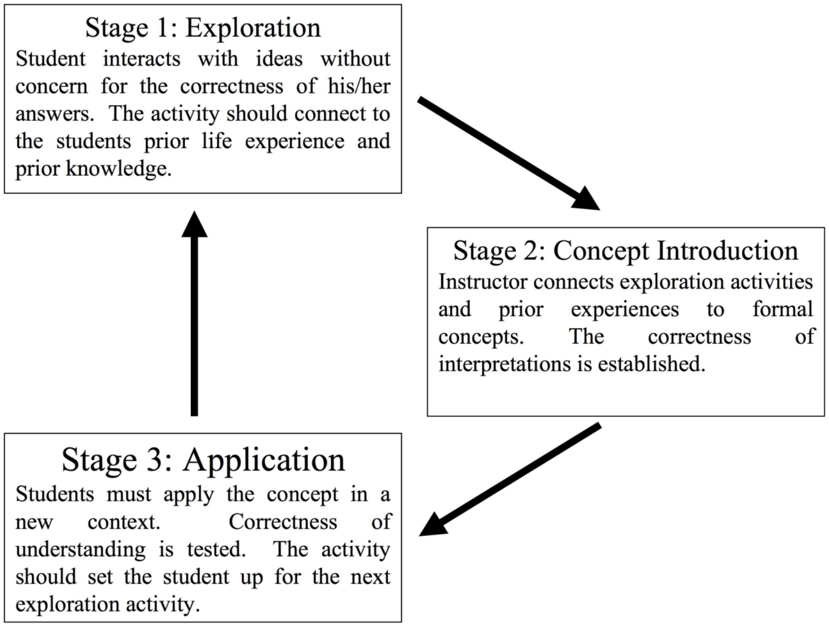
The social mechanism by which the ZPD is changed as described can then be termed scaffolding. Scaffolding, in analogy to its use in construction, is support provided by the more knowledgeable person to help the student construct sound knowledge (Wood, Bruner & Ross, 1976). Over time the idea is that the scaffolding will be removed, again in analogy to construction terminology, and concepts that were previously within the ZPD now reside within the student's current unaided ability. With this newly achieved competency the student will hopefully be ready to study material that was further beyond the ZPD, again with scaffolding provided by the more experienced teacher. Some research on the idea of self-scaffolding exists. This idea that the student may provide a form of socially interactive support for her or his own knowledge construction processes is an interesting one which might serve as an important conceptual bridge between Piagetian and Vygostky's constructivism.

The view that Piagetian constructivism and Vygotsky's theory of social learning are not in conflict, but address different aspects of the learning in a complimentary manner, is adopted here. At the same time, many other flavors of constructivist thought blend ideas from Piaget, Vygotsky and others. Readers interested in some of these different ideas might consider reading the article by Phillips for a broader discussion of different theories of constructivism in education (1995).

Vygotsky's work on social learning obviously relates to tutoring. In our work we must attempt to build our system to exploit a student's ZPD and use the SI-tutor as the knowledgeable expert. Since the system will not be intelligent, we will have to depend on the student to make choices that will ensure that they are working in their ZPD. This will be discussed in greater detail in chapter 3. From the opposite perspective, Vygotsky's work relates to ours in that future efforts using this type of system might set out to measure a student's ZPD. By building a tutoring system with tasks of differing difficulty and monitoring whether students can complete the tasks, and how much scaffolding is required to do so, the ZPD could feasibly be measured. Doing so with a system like this has benefits as compared to using teaching interviews. Since the computer-based system behaves the same way every time, it might provide a better standard of reference than human teacher-researchers. While the primary goal of this project is not to work towards this measurement, evidence for students ZPD will be noted for its value to future work.

2.1.3 The Learning Cycle

An important educational development that is consistent with a constructivist perspective is the learning cycle. The learning cycle is a way of organizing and presenting material that promotes the knowledge construction process (Atkin & Karplus, 1962; Karplus & Butts, 1978). Several types of learning cycles, differing mostly in the number of stages within the learning cycle, have been developed. The most basic is a three stage learning cycle, though five and seven stage learning cycles are also common. In the following paragraphs in this section the three stage learning cycle as implemented in this research will be discussed. The cycle is shown schematically in Figure 2.1.



Stage 1: Exploration
Student interacts with ideas without concern for the correctness of his/her answers. The activity should connect to the students prior life experience and prior knowledge.

Stage 2: Concept Introduction
Instructor connects exploration activities and prior experiences to formal concepts. The correctness of interpretations is established.

Stage 3: Application
Students must apply the concept in a new context. Correctness of understanding is tested. The activity should set the student up for the next exploration activity.

Figure 2.1 The three stage learning cycle

The first stage of our learning cycle is called the exploration stage. In this stage students perform an activity that is designed to help them build some kind of basic familiarity with the material that is to be learned. One of the important ideas of constructivism is that learning is built upon prior knowledge, and in many cases students will come to the classroom with a great deal of prior knowledge obtained by formal or informal means, but this may not always be the case. In the case that the student has a significant amount of prior knowledge the exploration phase serves to encourage the student to bring his or her attention to that knowledge and think about how these ideas are related to the material that is to be learned. It may also serve to encourage the student to explore related ideas that had not previously been explored and to develop and refine his or her prior knowledge. In the case that the student does not have much prior knowledge the exploration stage serves as a way of creating a base of prior-knowledge which can be built upon in the succeeding stages of the learning cycle. Exploration activities should typically be designed to be free from the constraints of right and wrong. The goal is not to develop knowledge that is consistent with an external authority of correctness, but for the students to freely create knowledge based on their experiences with a phenomenon. Since their experiences cannot be right or wrong, the knowledge created in the exploration activity cannot be right or wrong. This of course does not preclude the incorrect application of that knowledge in a subsequent activity, but within the exploration activity itself there should be a de-emphasis of the concepts of right and wrong.

The second stage of the learning cycle, called the concept introduction stage, is a more formal stage in which the events of the exploration activity can be discussed and connected to a formal introduction of the material that is to be learned. Evidence indicates that direct instruction can be an effective means of instruction if the student has some prior knowledge as well as some understanding about their difficulties with the material (Schwartz, 1998). Therefore, although concept introduction stages need not be direct instruction, such as a lecture, or reading a text, research indicates that this should not necessarily be viewed as an inferior method of constructing the lesson. The goal of the second stage of the learning cycle is to help the student generate organized knowledge, a schema, in which to interpret the events of the exploration activity and form a coherent understanding of the material being studied. This may involve showing the student how to generalize to fundamental principles from the specifics of what was observed in the exploration, or how the specifics of the exploration activity could have been predicted from general principles obtained by other means, or other types of discussion which serve to help construct deeper understanding.

The final stage is called the application stage of the learning cycle. This stage, in a sense, is about solidifying the newly built knowledge, with an eye towards formative assessment of student understanding. The goal of the application stage is to present the student with a new activity, which they have not seen before, and which requires the application of the material addressed in the first two stages. Successful application of this material has positive implications for the student's learning, and should also serve to further enforce the material in the student's mind. Unsuccessful application of the material has negative implications for the student's learning, but should serve as another exposure to the material, that should aid in the knowledge construction process, though further study may be needed for mastery. Once the learning cycle is completed the student should be in a position to start the next cycle and continue working through the larger curriculum. With this in mind, it is useful to construct the application activities in such a way as to connect to the next exploration activities, in order to take maximum advantage of what will soon be prior knowledge.

A learning cycle can be implemented on a number of time-scales, and with different numbers of students. The learning cycles developed for this proposed research are completed by individuals over the course of an hour or so, but learning cycles have been implemented in large lecture classes on the time-scale of a week (Zollman, 1990). The organizational and pedagogical benefits of learning cycles in education suggest that it would be beneficial to create the lesson materials that must be developed for this project in the form of learning cycles.

2.1.4 Ideas from Physics Education Research

Much work has been done by physicists to adopt and incorporate the findings and theories of cognitive science with experimental and observational studies of physics students into a coherent theory of learning physics. In this section we will give a brief review of literature relevant to establishing a coherent understanding of learning physics as a science. It has long been recognized that scholarly work focused on improving the teaching of physics was a worthwhile and perhaps necessary pursuit; this recognition resulted in the establishment of the American Association of Physics Teachers, which would be tasked with solving the problem of optimizing physics instruction (Richtmyer, 1933). Recognition that understanding how to best teach physics would have to be a slow, ongoing pursuit that drew on previous work in cognitive psychology was slower to follow, but a key development in physics education research (Mestre & Tougher, 1989; Reddish, 1994). This has resulted in a physics education research community that is generally receptive to the ideas from cognitive science, such as the constructivist ideas discussed previously.

Early efforts in physics education research focused on the study and remediation of common misconceptions in physics (Trowbridge & McDermott, 1980; 1981). These misconceptions can be robust against instructional intervention. Misconceptions research provides one classification scheme for student responses to conceptual questions. Therefore, familiarity with literature in this area can provide important insight into unifying themes in student responses to conceptual questions, such as the ones in our lesson materials.

A key idea that has emerged in physics education is that students' knowledge is typically fragmented (Hammer, 2000; Paul, diSessa & Roschelle, 1994). This view is in most ways consistent with the constructivist idea of knowledge building, but differs in that the term schema typically conveys a level of organization and coherence which need not be present in any given students knowledge. This has lead to the ideas of resources and resource activation in physics education. We can think of a resource as a knowledge element, which can be productively or unproductively activated in a given context (Hammer, 2000). In general, resources are not right or wrong, but usually context-dependent. A type of resource that is particularly important to this work is an epistemological frame. A student's epistemological frame addresses information about how they view the nature of knowledge within the current context (Scherr & Hammer, 2009). Clearly students can frame a learning situation in different ways, depending on how they view the way in which learning should occur (e.g. knowledge should come from an expert vs. knowledge should come from observation and experience). In our work we would like to foster an attitude that is grounded in ideas of knowledge is created through observation of nature and through social interaction with a more experienced guide (tutor).

Another concept important to physics education research is the idea of multiple representations: that a physics idea can be represented in multiple ways (Larkin, 1983; Dufresne, Gerace & Leonard, 1997). To recognize that this is true one need only realize that acceleration can be explained in words, equations, or depicted on a graph. Not all students are equally adept at understanding and using the different representations that we can associate with a given concept. Competency, however, demands a minimum level of skill across a range of representations, and this then becomes one measure of expertise. The idea that we can represent concepts in different ways, which may evoke different student difficulties, is important to our synthetic tutoring project because of the many ways we must convey information via our interface. The system will provide students with information via printed text, verbal/auditory information, video/pictorial information, and require them to interpret the information and provide printed responses, which may require mathematical manipulations to generate.

Each of these ideas that are commonly used in physics education research: misconceptions, resource activation, transfer and multiple representations are important to review for this research because they give us a variety of perspectives that we can adopt when looking at data from our synthetic tutoring system. We need to be familiar with a wide variety of perspectives so that we can see if and how they manifest themselves in our data.

2.2 Related Work on Human Tutoring

2.2.1 The "Two-sigma Problem"

In the early 1980's Benjamin Bloom and his students conducted seminal research on the effectiveness of one-on-one tutoring as compared to other instructional methods (Bloom, 1984). This frequently cited result is that one-on-one tutoring is far more effective than normal classroom instruction, or even mastery learning in which students studied material until they know the material well enough to obtain a minimum score on an assessment (Bloom, 1984). This is often characterized as the "two-sigma problem", that is students taught by tutor scored two standard deviations of the mean higher on assessments than students taught by normal classroom instruction and one standard deviation of the mean higher than students taught by mastery learning (Bloom, 1984). The "problem" in the "two sigma problem" is that one-on-one tutoring is not a very resource efficient method of instruction. For each student to be taught everything that they must know via one-on-one instruction is not feasible. Therefore Bloom set out to find methods of instruction that were as effective as one-on-one tutoring (Bloom, 1984).

Most of the methods that Bloom found that can rival the efficacy of one-on-one tutoring were composite methods, that is the combination of one or more instructional methods that did not yield such dramatic results by itself (Bloom, 1984). Ultimately, however most of these methods suffer from some level of complication in their application. One of the interesting facets of human tutoring is that it is so very effective, and so very simple. Bloom concludes his paper with a standing challenge to the educational community to continue to find instructional methods that have the potential to match human tutoring in efficacy (Bloom, 1984). Bloom's work provides great incentive to consider technological methods that were not available at the time of that study.

Beyond this well-known study, the efficacy of human one-on-one tutoring has been replicated and is well established. Cohen, Kulik and Kulik have done a meta-analysis of 65 studies on tutoring (1982). Of these studies 20 produced statistically significant results, 19 in favor of tutoring and 1 in favor of conventional instruction. The authors characterize the effectiveness of tutoring on the tutee in terms of effect size, which ranged from -1 to 2.3, with a mean of 0.400 and a standard error of 0.069 (Cohen et al., 1982). This mean effect size corresponds to a positive shift of 2/5 of a standard deviation (Cohen et al., 1982), which would be considered a modest effect by the standards established by statistician Jacob Cohen (no relation) (J. Cohen, 1977). Despite the mean effect being modest, commonalities were observed amongst studies that reported large effects. Identified traits shared by studies producing larger effects were:

1. more structured tutoring programs,
2. programs that were shorter in duration,
3. programs that taught lower level skills,
4. programs that focused on mathematics over reading,
5. programs that used locally developed assessments, and
6. programs that were published in journals (as opposed to dissertations).

Of these six commonalities, the last may be a self-selection effect associated with the difficulties in publishing less-spectacular results, the fifth may reflect either flawed or superior assessment, and the other four likely reflect adept use of tutoring in situations where it is in fact most effective. One-on-one tutoring implemented in an optimal manner will likely outperform typical implementations. Therefore, optimally implemented tutoring could outperform the mean effect size of 0.400. Beyond looking at efficacy, comparisons of students' attitude towards the material and self-concept were also performed. The authors conclude that tutoring has a statistically reliable and positive effect on students' attitudes towards a subject, but that any effects on self-concept are too small to be deemed statistically reliable. The authors point out that their results are consistent with other meta-analyses. Therefore a well-established literature documents the efficacy of tutoring as an instructional method (Rosenshine & Furst, 1969; Ellson, 1976; Fitz-Gibbon, 1977). Thus, we conclude that, even if Bloom's results are at the high end of what is realistically achievable, the superior efficacy of this instructional method is still reproducibly established, and a positive impact on students' perceptions of the material is established as well.

The question, then, is to what extent can a computer play a role similar to a tutor? This is one of the research questions outlined in Chapter 1 and is re-stated to remind the reader of the strong connections between Bloom's work and future work on developing computer/web-based instructional materials. As we will see in section 2.3 it is a question that is actively being investigated by researchers. One of the most useful pieces of information in investigating the possibilities associated with these technologies is what gives rise to the high level of efficacy. This will be discussed in the next section.

2.2.2 Explaining the Effectiveness of Human Tutoring

With the effectiveness of one-on-one tutoring established, the question still remains: why is tutoring so effective? Chi and collaborators have produced research towards answering this question. An intuitive explanation of the efficacy of tutoring is that it is just a matter of teacher attention. When an instructor must divide her or his attention amongst thirty students the quality of instruction will surely suffer. If instead the instructor can concentrate on just one student, then the instructor can focus effort towards understanding that one student's difficulties and tuning the instruction to address the student's individual problems. This has been termed the tutor-centered view of tutoring (Chi et al., 2001). There is however another perspective on the source of the benefits of tutoring. Perhaps something about being in the one-on-one learning environment enables the student to better construct their own knowledge, that is the benefits of tutoring are due to changes in student behavior, not changes in teacher behavior. This has been called the student-centered view of tutoring (Chi et al., 2001). The only other possibility, logically speaking, is that the benefits of tutoring are due to changes in the behavior of both teacher and student, and the interaction of these two changes in behavior. This is the interaction view of tutoring (Chi et al., 2001). Earlier research has suggested that in the domain of biology, student generation of self-explanation improves learning, as measured by recall questions, (Chi, De Leeuw, Chiu & LaVancher, 1994). This would point towards the student-centered view of tutoring, but is hardly conclusive. Later research provides more evidence in this direction. In this study eleven college students with superior content knowledge in biology, but little formal training in tutoring nor experience tutoring, tutored twenty-two eighth grade students in content related to the human circulatory system (Chi et al., 2004). The tutors worked with the students for a session, which was between an hour and a half and two hours (Chi et al., 2004). The study was naturalistic in the sense that the sessions were observed qualitatively in addition to the use of objective assessments (Chi et al., 2004). The observations were consistent with a picture of tutors who were inefficient in gauging student understanding and as a result inefficient in tuning their tutoring to meet students' needs (Chi et al., 2004). At the same time a burden was placed on the student to construct self-explanations, question those explanations, and refine them based on additional data (Chi et al., 2004). While this study cannot rule out the interaction view of tutoring, it does suggest a de-emphasis of the tutor's role in the efficacy of tutoring. This is a very interesting and counter-intuitive result, but potentially important for this project. It suggests that if our computer-based interface can encourage students to construct their own explanations, and challenge those explanations to make them fit all available information then it may encourage the same types of behaviors exhibited in the clinical study.

While Chi has attempted to ascertain the root of tutoring's effectiveness, Graesser and Person have examined tutoring from a more structural perspective; they propose a five-step collaborative model of tutoring (Graesser & Person, 1994). The five steps in this model of tutoring are:

1. Tutor asks a question or presents a problem,
2. Learner begins to answer question or solve problem,
3. Tutor provides short immediate feedback on progress,
4. Tutor and student collaboratively refine the solution or answer, and
5. Tutor assesses student's understanding of the answer.

Graesser and Person, as well as others have argued that tutoring is generally tutor-controlled, that is to say that the tutor determines the flow of session (1994). This is not in conflict with Chi and co-worker's observations of tutoring as a student-centered learning method (2001). The tutor may provide the structure for the session while the student may provide a minimum level of cognitive effort to progress the session. A similar perspective is advanced by Van Lehn and collaborators, who view tutoring in a way that is best described as coached-problem solving (2003). These researchers consider tutoring to be a goal-oriented activity consisting of a series of learning opportunities characterized by whether or not students reach an impasse in their problem solving activities, and whether it is the tutor or student who prompts a solution to the impasse. Again this perspective is not inconsistent with Chi and collaborator's work (2001). We can think of tutoring as a goal-oriented activity with the tutor directing the interaction while also thinking of it as a social activity which takes most of its instructional benefits from the pressure it places on students to build and evaluate new knowledge. In this work we consider these to be complimentary views of tutoring that are constructive for the development of our system.

2.3 Related Work on Human-Computer Tutoring

2.3.1 Computer-Human Coaching Research

Important work on the use of computers for physics instruction in a manner similar to human tutoring was done by Reif and Scott in the late 1990s. In that research a computer program called a Personal Assistant for Learning (PAL) was developed to help students with problem solving skills (Reif & Scott, 1999). The PAL system assisted students in learning problem solving skills in two ways (Reif & Scott, 1999). The PAL system could coach the student in solving problems by helping the student select the next logical step in a problem solving sequence, or be coached by the student as it worked through a problem solving sequence. Students working with the PAL would work with it in both modes of operation (Reif & Scott, 1999).

The research was conducted using volunteers from a large-enrollment physics course for science majors. Approximately 75 students volunteered from which 45 were selected to form three groups of 15 students, which were, based on SAT scores and prior academic performance, judged to be equivalent (Reif & Scott, 1999). The first group worked with the PAL system while the second group received human tutoring working on traditional problems similar to those used in the PAL system (Reif & Scott, 1999). The third control group participated in the class normally without any additional intervention (Reif & Scott, 1999).

The students in the first two groups worked on their respective physics assignments for five 90-minute treatments over the course of a week (Reif & Scott, 1999). A post-test was administered after the week of treatment. The PAL and tutoring groups scored comparably well, but both groups scored significantly better than the third group which had no intervention. The interesting result is that there was no statistically significant difference between the group receiving human aid and computerized aid. It is worth noting that the two groups who received the treatment not only did better than the normal class group at a level that was statistically significant, they also did better at a level that is of practical significance. The PAL group had a mean test score of 78.5±3.1; the tutoring group had a mean score of 84.0±3.5; and the normal class group had a mean score of 62.5±5.1 (Reif & Scott, 1999). This result, if reproducible has strong implications for the research proposed here. The PAL is a much simpler system than ours, but may already be able to approach the learning gains here, which it is worth noting, resemble the two-sigma data described by Bloom (1984). Research into the efficacy of PALs has continued into the present (Hsu & Heller, 2009). Reif & Scott's results and the community's continued interest in PALs give us strong encouragement to develop a still more interactive computerized learning environment.

2.3.2 A Social Component of Human-Computer Interaction

An interesting study which points to ideas that may relate to the research proposed here looked at the role of perceived interaction in human-computer interaction. In a recent study participants were told that they were being recruited and trained to tutor either a young girl via a virtual reality interface, or a computer system, again by the virtual reality interface (Okita, 2007). In reality there was no little girl, the tutors were always working with a computer system and the tutors' learning was in fact the subject of study. Participants who believed that they were interacting with a young girl retained more information than those who believed they were interacting with a computer interface. This work suggests some psychological tie between perception of interaction and learning, and while it initially looks as if the finding casts doubt on the utility of having students work with an interface that they know perfectly well is not a real human, there still remains much to be investigated in this research thread. A number of other permutations can be made based on who is being tutored, who is tutoring, and what they are led to believe about the nature of the interaction. These should be looked at, and although this is not the primary thrust of this research, the environment being developed may prove useful in further investigations of this type of question.

2.4 Multimedia & Digital Video in Physics Education and Beyond

2.4.1 Digital Video in Physics Education

The history of video's role in education is long, and it is not appropriate to outline it here. The important idea, in the context of this work, is that we can extract information about a physical system from video clips. Spatial information about an object is conveyed via its position within the frame, usually with respect to a reference object, or objects. Time information is conveyed via the inherently temporal nature of video. Each frame is of the video is separated from the previous and subsequent frames by the same time-interval: the inverse of the frame rate. This distinguishes the situation in physics education from the most common use of video in education: the produced educational film. No narration is needed to make the video useful, we need only look at the sequence of frames of video to understand what is happening in the physical system. We can think of the video clip as a form of data and the video camera that produced it as scientific apparatus; this is in fact what they are. Implementations of video of this type are relevant to our work.

The relevance of this type of video is clear when one considers the lessons that are a required component of our video system. Since our lessons are to be deployed on the Internet we are not restricted from using video, as we would be in print media. We are, however very much restricted in our ability to use real physical laboratory apparatus. We therefore have the opportunity to use short video clips, when appropriate, to enhance our lessons and in a very real sense, we must do that, because we cannot, due to the physical constraints of our system, connect the physical concepts to the real world of nature in any other dynamic way. The physical reality of video clips is not just an opportunity; it is very much a necessity. In this section, we therefore review prior work on video in physics education, video analysis in particular.

The directed study of interactive film or video clips in physics education goes back to the early 1970's. The Harvard Project Physics curriculum, which featured 8mm film loops used in classrooms to make quantitative measurements, is an early example (Holton, Rutheford & Watson 1971). The proliferation of relatively inexpensive technology for making and viewing film and video clips has enabled the use of this type of instructional technique. Today it is not uncommon for home video entertainment systems (i.e. DVD and to a lesser and lesser extent VHS players) and computer digital video players to have either frame-by-frame advancement or slow-motion capabilities. Since this early work several researchers have tested or advanced the use of video analysis (Noble, Zollman & Satern 1988; Zollman & Fuller 1994; Brungardt & Zollman, 1995; Beichner, 1996; Escalada & Zollman, 1997; Laws & Pfister, 1998; Brown & Cox, 2009).

Central to all of these studies, and video analysis in general, is the idea of getting time and distance information from a video clip and analyzing the information, either graphically or mathematically. Time information comes from the frame rate of the video. For example 0.033 seconds elapses between two frames of a 29.97 frame per second (a standard television or Internet video frame-rate) video clip. Counting frames enables the precise measurement of longer time intervals. Spatial information is obtained either from an on-screen length-scale, a calibration length in terms of pixels, or both. Early systems required students to mark out distances on acetate sheets attached to the video screen (Noble et al., 1988). More recently sophisticated user interfaces have been developed with built-in analysis capabilities. Many video encoding schemes interlace lines of pixels from adjacent frames of video. In order to extract information as described above it is necessary for the video to consist of non-interlaced frames. Video clips can either be shot with cameras which use formats that do not interlace the frames (this is commonly referred to as progressive scan), or for video clips that have been shot with interlaced frames it is possible to programmatically de-interlace the frames with a minimal loss of information.

Several studies have focused on the impact of real-time video analysis on graphing skills, in the domain of kinematics (Brungardt & Zollman, 1995; Beichner, 1996). Connecting graphs of kinematic quantities to physical models is a known area of difficulty for students (McDermott, Rosenquist, & van Zee, 1987). Work has also been done to incorporate video analysis into longer-term collaborative projects in mechanics (Laws & Pfister, 1998). The idea of a video analysis environment in which spreadsheet-style analysis tools can be brought to bear on video clips within the same environment as the video player have also been explored (Brown & Cox, 2009). These studies show the development in video analysis. In Zollman and colleagues' work in the late 1980's and early 1990's analysis had to be performed on acetate sheets physically attached to the screen (Noble et al., 1988; Brungardt & Zollman, 1995). The current state of the art in Brown and Cox's 2009 work allows students to click on an object in a video clip, log its position, and advance to the next frame. The software allows for graphs to be created automatically, or manually. While this type of system is powerful, it may present difficulties if a user wishes to do analysis with the software embedded in another system, such as a website. In that type of situation, which is relevant for our work, simple analysis within a stand-alone video player may be advisable. At present the QuickTime version 7 video player is one of the few (if not the only) embeddable media players that allows for frame-by-frame advancement. With the advent of YouTube, Flash video has become a household word; however the developments that make Flash video fast for streaming do not allow for the frame-by-frame capabilities this type of work demands. This information is important to anyone interested in doing video analysis.

The value of video analysis is sufficiently obvious to physics educators that commercial devices with no other purpose have been developed, and their educational utility explored

(Cadmus, 1990). Fortunately at this point, cameras and software are sufficiently common and inexpensive that instructors do not depend on these types of devices. At the same time, commercial software packages designed for video analysis, such as LoggerPro from Vernier, have also become available (http://www.vernier.com/products/software/lp, 2011).

Analysis of photographs has also been explored, and is another way to incorporate physical reality into multimedia instruction. Kanim and Subero, for example, explored a new application of an old technique by mounting flashing light emitting diodes (LED's) on objects and taking long exposure photographs of them during motion (Kanim & Subero, 2010). This work allows students to see, in a static photograph, the path of objects undergoing motion relevant to introductory physics such as constant velocity motion, constant accelerated motion, uniform circular motion, and simple harmonic motion. While long-exposure photography is a long-established technique in art and science, inexpensive digital cameras make this a practical technique for computer-based analysis. While efforts to assess the efficacy of these lab activities were minimal, and the results of those efforts were mixed, this work further demonstrates the ongoing interest in both qualitative and quantitative analysis of physical systems using multimedia.

Inexpensive digital video cameras, most recently web-cams, have also become widely available and allow teachers and students to make and analyze their own physics videos. Recent work directed towards understanding the fundamental uncertainties associated with these types of devices has shown that a proper understanding of the distortions and limitations of these devices can enable users to measure objects' distance traveled, or displacement, to within 0.1% (Page, Moreno, Candelas, & Belmar, 2008). For cameras with fast frame rate and short exposure time capabilities the uncertainty in derived quantities, such as velocity and acceleration, will be dominated by the uncertainty in displacement. This research therefore indicates that it is possible to use inexpensive web-cams to, under well-controlled circumstances, measure velocity or acceleration to similar precision. In a certain sense, this research takes video analysis out of the domain of exotic research-based pedagogy and places it well into the domain of common and inexpensive tools that are available for laboratory analysis.

A natural application of digital video is the combination of video clips and the Predict-Observe-Explain tasks mentioned in section 2.1.1. These tasks encourage students to compare their current understanding of a phenomenon with the phenomenon itself, and are consistent with the constructivist perspectives that ground our research. The combination of Predict-Observe-Explain tasks and video analysis does not appear to be widespread in the literature, but it is not without precedent (Kearney, 2004). As will be discussed in greater detail in chapter 3, video-driven Predict-Observe-Explain tasks can help us emulate the tutoring process, and so we will introduce them into our lesson materials.

2.4.2 Prior Research on Multimedia in Instruction

The idea of using multimedia to improve instruction is not new. The picture of research efforts to formally understand how best to use multimedia as an instructional tool however is hazy, complicated and somewhat lacking in cohesive focus. In this section we will describe some of the most relevant facets of the field of research.

Current computer technology gives us a high degree of freedom and control in creating instructional interfaces on the screen. We can distribute text, images, video/animations, and user controls (which naturally feedback to control the text, images, and video/animations) essentially however we choose. Tasked with using these capabilities to generate a multimedia interface that is optimized for learning, one might get a sense that there is too much freedom. Does it matter whether the text is on the right or the left? Does it matter whether this picture has text on it, or not? Should we use a video instead? The answers to these types of questions are not obvious. As researchers we might yearn for a coherent theory of multimedia instruction, which would provide experimentally verified principles of design that can be counted on, and taken into account when designing multimedia-based learning interfaces. Cognitive Theory of Multimedia Learning (CTML) represents an effort to achieve this type of theory (Mayer, 2005). CTML is heavily informed by the ideas of Cognitive Load Theory (CLT), which centers on the amount of cognitive effort required to perform tasks, and how that effort increases or decreases as the task is changed in various ways (Sweller, 2005). Central to CLT is the idea of working memory, and its finite limit. A good functional definition of working memory is the quantity of concepts that a person can concurrently work with at a given time. One of the most famous results in cognitive psychology puts this at 7±2 (Miller, 1956). For our purposes the exact number is not nearly so important as the fact that it is finite and small. An intuitively obvious facet of working memory is that we can collapse elements together to make them easier to remember. Most of us would have trouble remembering a string of ten numbers, but few of us have trouble remembering someone's telephone number, if it is required. By turning ten elements into one we have turned a hard task into an easier one. This was idea of chunking was apparent to Miller in the same research conducted in the 1950's. CTML is based on the idea that different instructional multimedia interventions place different amounts of cognitive load on the student and that this cognitive load can be a barrier to learning. By experimentally varying different components of multimedia instructional interventions researchers seek to identify design principles that reduce cognitive load. Researchers in CTML believe that it is possible to optimize multimedia instruction using these design principles. Many basic design principles have been proposed, primarily by Richard Mayer, mostly as a result of experiments involving problem-solving transfer tests administered shortly after instruction. Readers interested in these design principles are encouraged to reference The Handbook of Multimedia Learning edited by Mayer (2005).

Several issues arise directly when discussing CTML, the first being the generality of these principles. For example, research on the teaching of physics by Stelzer and collaborators in which a plain text instructional intervention was more effective than an intervention which used images and words (Stelzer, Gladding, Mester & Brookes, 2009) calls even the first principle into question. In this work researchers compared the efficacy of a plain text instructional intervention, an intervention based on a popular, modern (and carefully designed) illustrated textbook, and a dynamic multimedia intervention, finding the multimedia intervention most effective followed by the plain text intervention and then the textbook centered intervention (Stelzer et al., 2009).

A second issue arises out of choosing to base CTML on CLT, and the difficulties in characterizing what exactly working memory is, and quantifying it. As has already been discussed, ideas can be collapsed into organizational schemes that "free up" working memory for other tasks. It is likely that different people will do this in different ways. Ultimately many ways exist to think about multimedia instruction. Ascertaining which one(s) best contribute to the various facets of multimedia learning is a challenging and on-going process.

Therefore because of these difficulties and ambiguities, and despite the fact that this area of research is very much related to our research goals, we believe that our research is far more likely to contribute to the development of a better understanding of best practices of instruction with multimedia than it is that existing theories of instruction with multimedia will provide the answers to questions of how to best design our system. Several of the principles that are advanced by CTML researchers do seem intuitive and we certainly designed our system to reflect those principles. However, we believe that by intuitively working to build a parsimonious instructional system we likely achieved results that are as good, or better than what we might have achieved by working from these principles.

Chapter 3 - Research Design and Methodology

This chapter focuses on research methods needed to evaluate the efficacy of our synthetic tutoring system, and which will hopefully enable us to learn more about how people learn physics in general. It begins with a discussion of the population of interest. This is followed by a detailed description of the system that we are developing for this research project: the Pathway Active Learning Environment. In this chapter we will discuss its role both as a subject of study itself and also as a tool for studying student understanding and learning. The proposed research designs are then discussed. This includes discussions of methods of obtaining participants and data collection methods. Potential data analysis schemes will also be discussed. This chapter will conclude with a review and summary of the research for the dissertation.

3.1 Populations of Interest

In conducting this research we are targeting students enrolled in a typical high school physics class (not an Advanced Placement (AP) physics course) and college students enrolled in algebra-based or concept-based introductory physics courses. While it is not advisable to consider these three groups of students to be interchangeable, members of these three groups do share many common characteristics. Furthermore there is a great deal of commonality in the physics that they are studying. We therefore believe that it is possible to develop an interactive learning environment that will be useful to all of these populations and to the extent that it is not possible to serve all three populations equally well, the high school students will be favored, keeping with the historical theme of the project: improving physics education in the high school environment.

Beyond historical reasons, our choice of target population is quite logical. High school and college students of introductory physics by far account for the vast majority of students enrolled in physics courses. The student to teacher ratio for a high school physics class can easily reach 30 to one, but when taking into account a teacher's full teaching load the ratio can soar over a hundred to one. A large-enrollment college physics class has a student to teacher ratio that starts at a hundred to one but can reach as high as a thousand to one at the nation's largest universities. Recitations provide slightly smaller forums for discussion, but an instructor teaching two or three recitation sections will again be responsible for teaching about a hundred students. One-on-one time with an instructor in any of these classes is a rare commodity. This harsh truth of large-scale education is particularly troublesome because it impacts the students who are least familiar with the material, and therefore least able to help themselves. Advanced students, in classes with student to teacher ratios that can be as low as ten to one but are rarely higher than 30 to one, are far more self-reliant and sophisticated in their study habits. Our choice of populations make sense because they are composed of students who are likely to be most in need of one-on-one assistance, and for whom that assistance is most difficult to provide. In targeting these populations we are also setting ourselves up to have maximum educational impact.

3.2 The Pathway Active Learning Environment

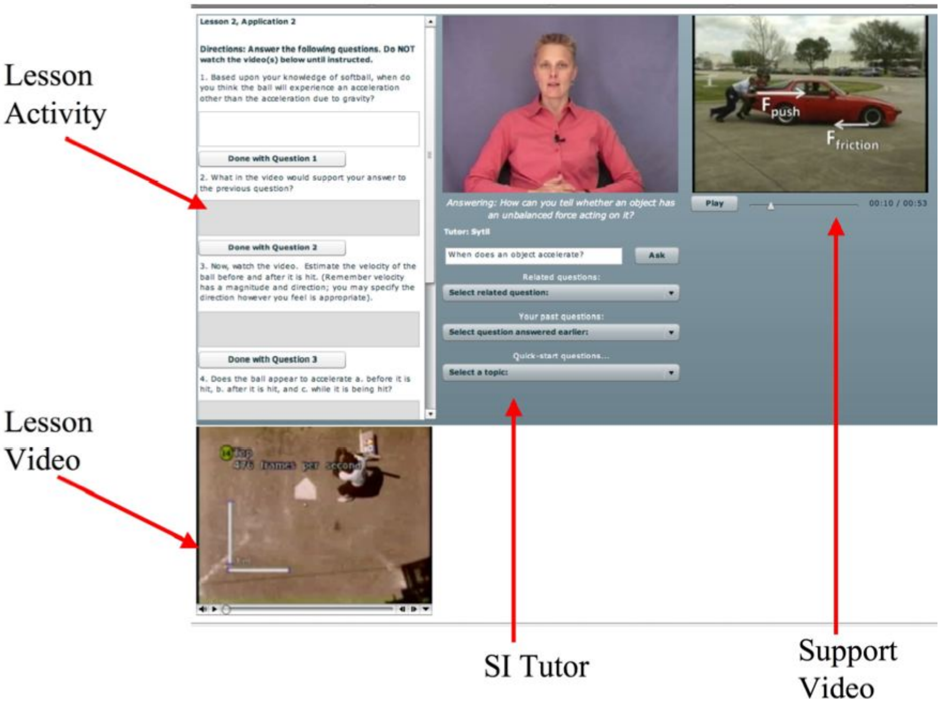
The Pathway Active Learning Environment (PALE) is both a tool for probing student learning and the subject of study in this proposed research. At its core it is comprised of three components, each of which will be discussed in detail.

The heart of the PALE is the Synthetic Interview (SI) tutor that allows students to ask natural language questions and get pre-recorded video responses. This component makes it possible for the PALE to progress from a collection of activities to potentially become a tutoring interface. The SI interface is discussed in detail in section 3.2.1.

The second key component of the PALE system is a set of lesson materials that enables the tutoring system to have a structured flow. Most tutoring occurs in the context of some sort of homework assignment, or problem set. Without that context what we wanted to be a tutoring session is in fact nothing more than a Q&A session. Beyond providing a context in which to construct a tutoring interaction, we have tried to construct our lessons activities to connect to accessible real-world physical scenarios, with each lesson containing one or more video clips that establish that connection.

As discussed previously, video analysis is a key facet of these lesson materials. To promote video analysis we need a video player that allows frame-by-frame advancement, and video clips with non-interlaced frames. To this end we have adopted QuickTime as the player for the lesson videos. Homemade video clips were shot at 30 frames per second with a camera designed to be capable of shooting in a non-interlaced mode. Video clips which were not homemade were tested prior to use by students to ensure that they behaved properly when advanced frame-by-frame.

The final component of the PALE is support of multimedia which enables the SI tutor to provide relevant, compelling examples that clarify and explicate the physics being discussed. A real tutor will draw pictures, sketch motion or provide other visual support when teaching a student, and the support multimedia allows our synthetic tutor to provide visual aids that serve a similar purpose. A screen capture of the synthesis of these three components into a practical user interface is shown in figure 3.1



Lesson 2, Application 2

Directions: Answer the following questions. Do NOT watch the video(s) below until instructed.

1. Based upon your knowledge of softball, when do you think the ball will experience an acceleration other than the acceleration due to gravity?

2. What in the video would support your answer to the previous question?

3. Now, watch the video. Estimate the velocity of the ball before and after it is hit. (Remember velocity has a magnitude and direction; you may specify the direction however you feel is appropriate).

4. Does the ball appear to accelerate a. before it is hit, b. after it is hit, and c. while it is being hit?

Answering: How can you tell whether an object has an unbalanced force acting on it?

Tutor: Sytil

When does an object accelerate?

Figure 3.1 The Pathway Active Learning Environment interface

3.2.1 The Synthetic Interview Tutor

The SI technology was developed by collaborators at Carnegie Mellon University to enable people to interact with a computer in a more social manner (Stevens et al., 2007). The basic idea behind the SI is that a computer algorithm, which is discussed in greater detail in section 3.2.1.2, can be used to match up a question asked by a user to an element in a master list of questions. This matched item should ideally contain the same, or similar, information to the question asked by the user. The computer algorithm then retrieves a pre-recorded video response to the matched item. A better match between the item in the master list of questions and the question asked by the user typically means a more satisfactory match between the student's question and the SI tutor's answer. This is shown schematically in figure 3.2.

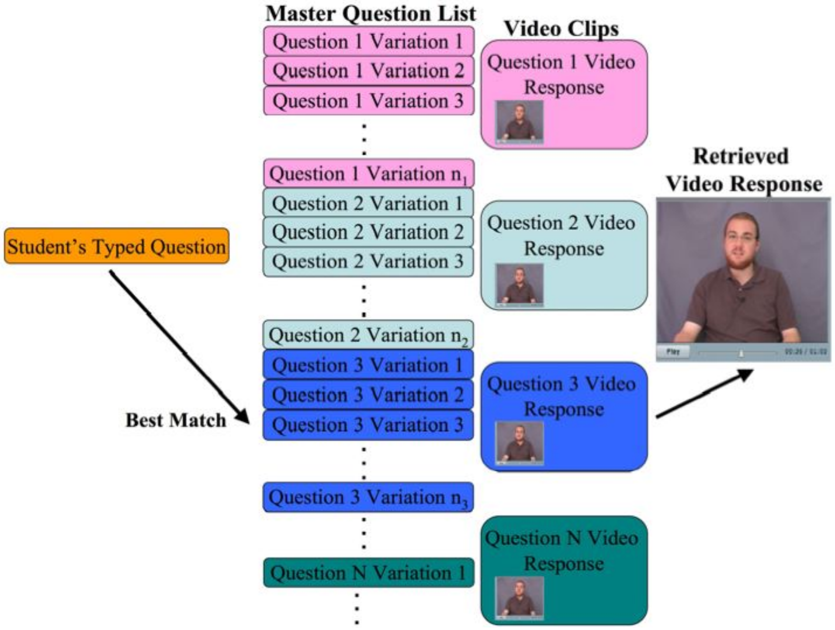


Figure 3.2 Schematic of SI-Tutor's functionality

In order to create this SI interface we need three components. We need a list of conceptually distinct questions which cover the range of information the SI will be able to provide, a set of pre-recorded video files which provide the SI's response set, and an extended list of questions which includes variations on the conceptually distinct core questions. As will be discussed in section 3.2.3, we in fact need several sets of SI video responses to allow the system to answer questions with and without additional multimedia support.

In order to generate the first question list, which spans the content and determines what video responses are needed, we have conducted interviews with several students from the target populations. In the interviews the students worked with paper and pencil versions of some of our lesson materials (described in detail in section 3.2.2). These experiences provided some typical questions that students would ask while working on these types of materials. We then drew on the teaching experience of several participants within the collaboration to extend this question list and cover the full range of relevant material. Obviously some judgment was required, and we do not currently have an established protocol that will generate a "best list" of questions, if such a thing exists. Establishing effective and efficient methods of producing this type of question list would be a valuable result of this work, if we can obtain it.

Generating the set of video responses is, in principle, as simple as selecting a person to answer the questions, presenting them with the questions and putting them in front of a video camera. In practice, however important factors must be considered when building our bank of video responses. If we are to create a pedagogically useful tutoring system we recognize the importance of building the system around SI tutors that students can relate to and identify with. Because of the variance in the student populations that we would like to study, we should also build variance into the SI tutoring system. To that end we recruited three people to record video responses. All three are experienced teachers of physics, who are either pursuing or have achieved the PhD degree. At the same time all three are also relatively young, and therefore may better connect with a student population. They are shown in figure 3.3. We believe that selecting multiple expert tutors will allow us to give students multiple perspectives on the physics and produce a more useful tutoring system. Unfortunately due to scheduling constraints on the part of one of the experts, Nasser Juma, we were unable to record all of the video responses necessary to incorporate all three experts into the tutoring system on a time scale that was reasonable for this dissertation. The other two tutors recorded all the necessary video responses. We hope that in the future we may have the opportunity to again work with Mr. Juma, and to work with other expert tutors, to increase and improve our system's offerings.

Figure 3.3 Tutors recruited for the project

Building the extended question list, much like the core question list, requires a fair amount of judgment, and again we do not have a well-established protocol for generating an optimized list. In order for the system to work properly the extended question list must have a certain amount of redundancy built into it. If there are one hundred pre-recorded video responses corresponding to one hundred conceptually distinct questions, there must be far more than one hundred questions in the extended list to account for variations in how different users might ask the same question. For example "What is the relationship between force and acceleration?" and "How are force and acceleration conceptually connected?" are two valid ways to ask the same question. Clearly if one had to catalogue all possible ways of asking each question, then one hundred questions could balloon into tens or hundreds of thousands of questions very quickly. In fact, finding all of the ways of asking a given question or even most of the ways is not necessary. Our initial experience building the system suggests that just fifteen or twenty variations on the question may be sufficient to provide a good match most of the time. Some trial and error is required to generate an extended question list that works well for tutoring students in kinematics and Newton's laws.

3.2.2 The Pathway Active Learning Environment Lesson Materials

A tutoring session does not typically consist of a simple exchange of questions and answers. Instead, a task, such as homework, typically prompts the student to seek the tutor's help. This task provides structure and direction for the tutoring session and if the PALE is to actually emulate tutoring in any way, it must have a similar component. To that end a series of lessons based on the three-stage learning cycle discussed previously were developed. Early test versions of these lessons can be viewed at: http://perg.phys.ksu.edu/altpathway. The lessons center on Newton's laws, with one lesson written for each of the laws, but by necessity draw heavily on subject matter from kinematics. Having developed these lessons developing additional lessons which address kinematics topics directly would not be too difficult. In particular because of the focus on the high school physics student population, this may be desirable because of the large amount of time high school physics courses frequently dedicate to kinematics. Kinematics is also, in many ways, pre-requisite knowledge for further studies in physics.

All lessons focus on using video clips to encourage students to make observations, measurements and inferences about real phenomena, sometimes in natural contexts, sometimes in contrived contexts. Students are able to respond to the questions in these lesson materials via some form of online response system, like a survey system, but paper and pencil worksheets may be appropriate in some instances. In the next three sections each of the lessons will be described in detail. All the lesson questions are included in Appendix A of this dissertation.

3.2.2.1 Lesson 0: Learning to use the system

Before using the PALE system, we believe that it is helpful, perhaps even necessary for students to spend a little bit of time learning how the PALE works in a context that is very low in cognitive demand. To accomplish this goal a "zeroth lesson" was created which directs students to use the PALE to better understand its functionality.

The lesson consists of one activity that breaks up neatly into two sections. The sections are driven by lesson questions, just as they will be in the lessons on physics. The first section consists of questions related to aspects of the SI tutor's personal life that are socially appropriate to ask about in a teacher-student context. The students are instructed to find out different pieces of information about their tutor (e.g. alma matter, hobbies, pets, etc…) that can only be obtained by asking the tutor questions. In this way we will attempt to show students that asking the tutor questions is a productive action.

The second section of the zeroth lesson focuses on the idea that we can use the frame-by-frame capabilities of the QuickTime player to extract useful information about objects in a video clip. A video of a cart rolling across the screen is presented, with an onscreen scale present. The students are asked to ascertain how many frames it takes for the cart to cross the screen, what that corresponds to as a time interval in seconds (a frame rate of 30 fps is provided), how far the cart has traveled in that time interval, and how fast the cart is moving as it crosses the screen. This section of the lesson is designed to help students get used to extracting information from the QuickTime videos in a very simple context that is extremely low stakes. At the same time, in the course of the physics lessons the students would rarely have to use the video clip in a more complicated manner than required in this introduction. We therefore believe that it is quite reasonable to presume that a student who can extract this information can reasonably extract the information needed for other lesson activities, assuming they understand the context of those activities.

Research designs are discussed in greater detail in section 3.3, but at this point in the discussion it is logical to note that it was desirable to develop our system with the option of allowing students to work with our lesson materials without the SI tutor, this ability would allow us to get an idea of the effect of the lessons alone on learning, and from this we can infer something about the effect of the SI tutor. Students who work through the lesson materials without the SI tutor would not benefit from the first portion of the introductory lesson, but would still benefit from the second. We therefore designed the system so that students who work with the system in a lesson-only mode are given the second portion of this introductory lesson and not the first.

3.2.2.2 Lesson 1: Newton's first law

Once students have had some experience working with the system and understand the basic interactions that are available they can move on and begin studying physics. The lesson, like the remaining two, contains six video activities: three activities in the exploration section and three activities in the application section. To explain the activities in detail while eliminating excessively verbose passages of text, tables have been created to explain each section of the lessons. The lesson 1 exploration activities are discussed in Table 3.1. The lesson 1 application activities are discussed in Table 3.2

3.2.2.3 Lesson 2: Newton's second law

Similarly to Lesson 1, the lesson 2 exploration activities are discussed in Table 3.3 and the lesson 2 application activities are discussed in Table 3.4.

3.2.2.4 Lesson 3: Newton's third law

Similarly to lesson 1 and lesson 2, the lesson 3 exploration activities are discussed in Table 3.5 and the lesson 3 application activities are discussed in Table 3.6.

Table 3.1 Lesson one exploration activities

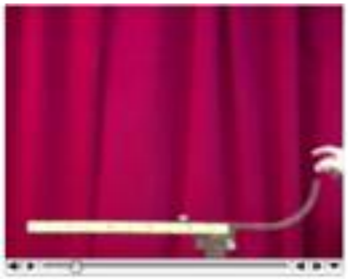
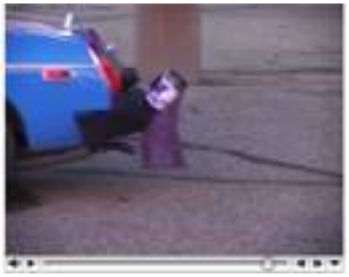
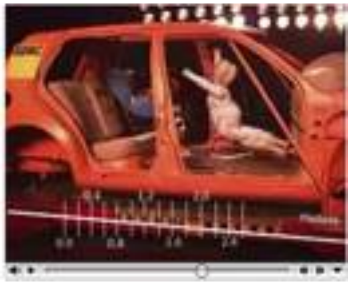
Exploration	Image	Activity Description	Central Idea
Activity 1		Students measure the speed of a ball rolling on a flat, smooth track. The speed is measured near the beginning and end. Students are asked to consider the expected speed in the middle of the track and at the end of a longer track.	Students should find that the speed of the ball does not change significantly from the beginning to the end. They should then infer that the speed should be the same in the middle, and at the end of a longer track. Addresses the idea of inertia in the context of a moving object.
Activity 2		Students are asked to consider the case of a coffee cup "accidentally" left on a car. When the car moves, how does the cup behave?	A painted line in the video clip provides evidence that the cup tends to stay in one place, at least horizontally. This addresses the idea of inertia in the context of a stationary object.
Activity 3		Students consider the case of a crash test dummy that is unrestrained during a crash test. An on-screen scale allows them to estimate the speed of the car and dummy before and after the crash.	This again gets at the idea of inertia in a moving context. It is not unusual for people to believe a force throws someone from a moving car, when in fact it is more typical for their inertia, and present velocity to account for the behavior.

Table 3.2 Lesson one application activities

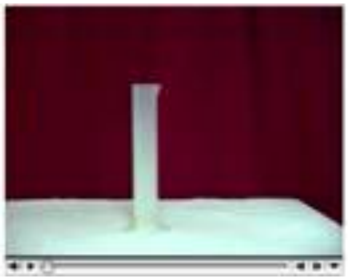
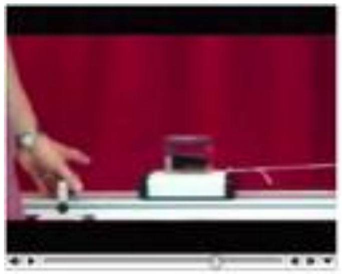
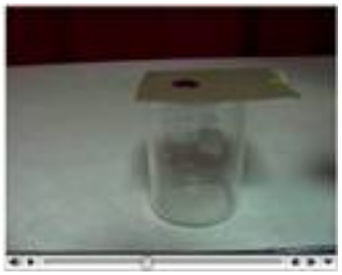
Application	Image	Activity Description	Central Idea
Activity 1		Before watching the video students are asked to explain, using Newton's 1st law, how to remove a coin that is stuck in a graduated cylinder. Afterwards they are asked to explain whether they were right, and why the demonstrated method works.	This activity is very similar to the car crash activity. The cylinder and coin are set into motion together, but when the cylinder is brought to a rest, by the table, the coin continues to move and can be obtained.
Activity 2		Using their knowledge of Newton's first law students are asked to predict the profile of surface of water in a container on a cart which is accelerated by a constant pull. They are then asked to determine whether their prediction is correct, and why the surface looks the way it does.	The surface of the water slants upwards towards the back of the cart. This connects to Newton's first law in the sense that the fluid, which is massive, resists changes in motion. It is similar to the car and coffee cup experiment.
Activity 3		Students are asked to predict the trajectory of the coin, when the card is pulled rapidly. They are then asked to explain the result.	Like the car and coffee cup video this addresses inertia in a static context. The coin falls straight down. The frame-by-frame feature of the video player makes this easily verifiable.

Table 3.3 Lesson two exploration activities

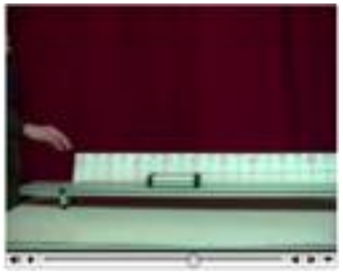
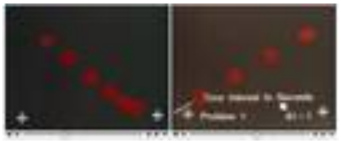
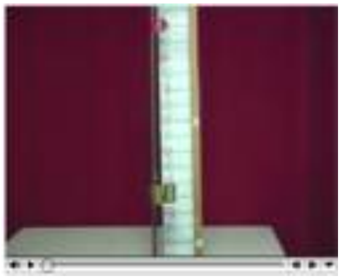
Exploration	Image	Activity Description	Central Idea
Activity 1		Students measure the acceleration of carts in three conditions. The first case is a cart of mass 1 unit, pulled by a force of 1 unit. The second is a cart of mass 1 unit pulled by 2 units of force. The third case is a cart of mass 2 units and a force of 2 units. They are asked, what is the simplest relationship that explains this data?	This activity shows that the acceleration is proportional to the applied force, and inversely proportional to the mass of the cart. This is both the simplest relationship that could explain the data and Newton's second law.
Activity 2		Two videos are used to contrast constant acceleration and momentary acceleration. In the left video the puck has a constant acceleration and its displacement over equal time intervals increases. In the right video the puck accelerates only when it is set into motion. Its displacement is the same over equal time intervals	The central idea is that an object only accelerates while a non-zero net force acts on it and it is accelerating only when a non-zero net force acts on it. This is in contrast to the belief that the force must act to keep the object in motion.
Activity 3		This video contrasts three cases: two Attwood machine set-ups with equal mass on each side, and one Attwood machine set-up with unequal mass.	We address the idea that the net force is the sum of applied forces. When the masses are equal, no net force acts because the force due to the tension cancels the weight. In the other situation the net force is non-zero and the objects accelerate, but not at 9.8m/s^2, because the force due to the tension cancels some of the weight.

Table 3.4 Lesson two application activities

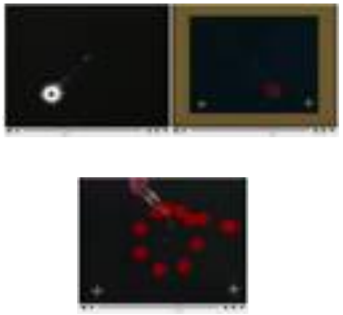
Application	Image	Activity Description	Central Idea
Activity 1		Students apply their understanding of Newton's second law in the context of an experiment performed on the moon: the dropping of a hammer and feather. Since both fall at the same rate, their respective accelerations are equal, but each feels a different net force, proportional to its mass.	Given plausible mass and acceleration values students should be able to calculate the net force on each object and explain what is occurring.
Activity 2		Students can estimate the magnitude and direction of the net force on a softball when it is hit by a bat. This is done by estimating the incident and receding velocities and the time over which the bat and ball are in contact in order to find the acceleration. The mass of a softball is given.	This is a straight-forward application of Newton's second law, although the fact that the ball changes its direction of motion provides a potential complication.
Activity 3		Students are asked to recognize the vector nature of force, velocity and acceleration by realizing that if an object's velocity is changing direction it must be accelerating and therefore must feel a non-zero net force.	This activity connects Newton's second law to circular motion and stresses the vector nature of this relationship.

Table 3.5 Lesson three exploration activities

Exploration	Image	Activity Description	Central Idea
Activity 1		Given the initial speed for each train, and their (identical) masses students can estimate the acceleration, and net force on each train as they come to a stop.	The net force on each train is clearly due to the other train. Since their masses, initial speeds, and final speeds are also the same it is easy to show (and understand) that the forces are equal and opposite. This activity sets the stage for unequal masses and speeds.
Activity 2		Given the mass of a bowling ball and an ice skater, as well as the knowledge that they are initially at rest students explore the force on each when the ice skater throws the ball.	Students should find that the bowling ball and the ice skater feel equal and opposite forces, demonstrating that Newton's third law holds even when the masses are unequal and the objects start from rest.
Activity 3		Students examine collisions between two carts. The incident cart is in motion, while the target cart is stationary. Students can use the frame-by-frame capabilities of the video player to measure the initial and final velocities for the two carts.	This is a slightly more complex situation in which one object is stationary and the other is moving. Again they should find that the forces on the respective carts are equal and opposite.

Table 3.6 Lesson three application activities

Application	Image	Activity Description	Central Idea
Activity 1		Students compare two carts which have propellers mounted on them. The first cart has only a propeller, and it accelerates when the propeller is turned on. The second has a propeller pointed at a sail mounted on the cart. It doesn't accelerate when the propeller is turned on. Students are asked to explain this behavior using Newton's third law.	In case one, the blades exert a force on nearby air, which exerts an equal and opposite force on the blades, and the cart, causing acceleration. In the second case the blades exert a force on the air, and the air on the sail but the sail exerts an equal force back on the air, and the air an equal force back on the blades. The net force on the cart is zero: no acceleration.
Activity 2		In this activity a car crashes into a wall. The car obviously accelerates as it comes to a rest, and therefore feels a non-zero net force. The wall doesn't accelerate, however. Students are asked to reconcile this fact with Newton's third law.	The wall is likely fixed to the floor of the building, which is fixed to the Earth. Therefore although the wall does feel an equal and opposite force the effective mass of the wall is so large that a zero acceleration situation is believable.
Activity 3		In this activity two balls of equal mass, but different mechanical properties are dropped on a table surface. One ball bounces and one ball falls flat. Students are asked to figure out which ball is subjected to a greater force and which ball exerts a greater force on the table, or you if it were dropped on you.	The key idea here is that because of the vector nature of velocity the ball that changes direction (bounces) accelerates more than the ball that falls flat (assuming a fixed time interval, which is fair to assume). Therefore the bouncy ball feels more force. Consistent with Newton's third law, that ball must also exert more force.

3.2.3 Pathway Active Learning Environment Support Multimedia

The PALE should be capable of supporting the tutors' video responses with image or videos that further illustrate, clarify, or exemplify the topic of discussion. One of the questions that we addressed is how support media can help or hinder learning. We therefore built the PALE system with multimedia support in mind. This could not be implemented independently from the SI tutor. A real tutor's response changes depending on whether s/he has some picture or demonstration to reference. No one will draw a picture and then not reference the drawing. Allowing for multimedia support clearly necessitates recording separate SI video responses for all questions that are meant to have multimedia support available.

Two natural modes of multimedia support should be considered. These are static images, and dynamic videos. In a real one-on-one tutoring session a tutor is unlikely to provide video examples to clarify the topic of discussion (though, with the advent of streaming video, not impossible). In this regard our system is more flexible than a human. At the same time, a video is not obviously better than a picture. If a picture is worth a thousand words, how many words is a video worth, and more importantly, how many words is too many? A key region of interest in this project is to investigate whether students prefer one of these modes of presentation and whether one of them could be more effective for instruction. Neither is likely to emerge as universally more liked and more effective. We therefore cast this question as an exploratory one, in which we used interview data to explore the circumstances that may give rise to one being preferred over the other.

An important factor to consider in designing the support media is that we can represent physical ideas in many ways and different students will perceive different representations to be more accessible. As we developed the PALE system we gave consideration to different methods of representing physics concepts within this support media.

3.3 Research Designs and Data Collection Methods

Because of the large number of potentially interesting research questions that may be available to this line of research, and also because it is not necessarily possible to determine a priori which research questions will emerge as the most interesting, or the most accessible it is important to develop a flexible research design which allows for the collection of a wide range of broadly-usable data. Doing so allowed for the extraction of information in a more exploratory manner that revealed which research questions are interesting and accessible.

The simplest research design is a two group experimental design in which student learning is compared for students who worked with the PALE and for students that did not. This has an advantage in that given enough resolving power and the right kind of assessments, the design can unambiguously determine whether the PALE presents learning advantages and measures the size of those learning advantages. The problem with this type of design is that, at best, it only will yield a yes or no answer to the question of whether the PALE is more effective than a control. Of greater concern is that, in the “no significant difference” result, we are left with no real information about what the system did do. We know that the two interventions are not identical, and thus, the students in the two groups did not have identical experiences. However, we only have a result that casts them as identical, as far as we can tell. We need a research design that is more nuanced, and that gives us a picture of how students experience the system, even if the system does not produce learning gains that are superior to traditional instruction, as measured by pre-test and post-test. Therefore the design which we implemented is one in which student responses to objective questions, subjective questions, and a detailed picture of their interactions with the PALE are put together to generate a complete picture of how the PALE impacts student learning, and how students understanding may change over the course of their interaction with the environment.

3.4 Sampling Methods

We studied student use of the PALE system with several student populations:

1. High school physics students in non-Advanced Placement (AP) physics courses,
2. College physics students studying physics from a largely conceptual perspective, and
3. College physics students studying physics using only algebraic or trigonometric math.

When considering sampling methods it is important to consider the issue of access. We conducted our research in ways that are minimally invasive and maximally constructive for the students. In order to study our system our sampling methods focused on soliciting volunteer teachers who were interested in using the system with their students. We have had contact with several high school teachers who expressed interest in trying the system with their students in-class. We contacted an algebra-based physics instructor who gave us permission to solicit volunteers from his class for interview studies. We also contacted a concept-based physics instructor who was willing to make our materials part of his class' assigned homework. Between these three sources of students we believe we have enough students to study the functionality of the system. While the system is ultimately designed to be used at home for supplemental instruction, it is difficult to study the workings of the system in that context. We therefore studied the system in multiple research contexts: at home, in classrooms and in our one-on-one interview setting. The number of students recruited in each student population, the context in which we studied their use of the system and the semester in which the data was collected is outlined in Table 3.7

Table 3.7 Student sampling for PALE studies

Student Population	Time frame	Research Context	Number of Students
Algebra-based physics students	Fall 2010	1:1 interviews	22
Algebra-based physics students	Summer 2011	Classroom use	30
High school physics students	Fall of 2010	Classroom use	41
Concept-based physics students	Fall of 2010	At home use	99

Demographic information was collected for all the college students who used the system. The gender distribution for the algebra-based physics students who participated in interviews was 14 female and 8 male. All but one of these students identified as White for their race/ethnicity. The remaining student identified as "other". The gender distribution for the algebra-based students who used the system in the classroom context was 18 female and 12 male. Twenty-three of these students identified as White, 2 identified as multiple races, 2 as Asian, 1 as Black, 1 as "other, and 1 declined to provide race information. Only 70 of the 99 concept-based physics students submitted demographic information. Of these students, 65 (92%) were female. Fifty-six of these students (80%) identified as White, 6 (9%) identified as "other", 5 (7%) identified as multi-racial, 2 students declined to provide race information and 1 student identified as Asian.

The Consistent with Institutional Review Board (IRB) standards all interview research volunteers were given informed consent forms which inform participants of their rights, the nature of the research project and the data collection protocols they are being asked to help us with. In our research students were also informed about the nature of the research via a terms of participation screen which was displayed on the website itself.

3.5 Data Analysis Methods

Since we have several types of data we applied several types of analysis techniques to extract a useful picture of how students might constructively use the system, as well as how to improve its functionality. Here we present a three-part analysis. The first part is a qualitative, analysis of transcripts of interviews with algebra-based physics students. The second part is a descriptive analysis of the data from the PALE log. The third stage of analysis combines automated text classification algorithms with our ability to better interpret student meaning into a process that enables automatic adaptive feedback to students. We believe that this third analysis protocol will provide focused answers on whether or not this type of system can really begin to emulate the process of tutoring.

3.5.1 Qualitative Analysis of Interview Data

In order to understand facets of students PALE use that do not translate well into PALE log data we conducted 59 interviews with 22 algebra-based physics students who had volunteered to use the system. The design is described in detail in section 3.3. In order to make sense of students' experiences with the PALE we transcribed the interviews, and coded the transcripts sentence by sentence looking for themes that emerge in their descriptions and comments. While we paid attention to all themes that emerge, we are particularly interested in statements of several types. We are interested in statements that:

1. give evidence that students might or might not be willing to engage with the SI-tutor in a way that we might consider socially interactive,
2. give some evidence that students would or would not use this system in their studies (in the absence of external reward beyond potential learning benefits),
3. show what features the student does or does not like about the system,
4. gives some insight into how the student used the system or framed the activities, and
5. give evidence about students physics knowledge beyond what might be logged in the PALE log.

The protocol was designed to elicit responses that focused on these types of themes, but at the same time, was semi-structured to allow freedom for the interviewer to probe into whatever topic the student is inclined to discuss.

Transcribed interview data were analyzed will from a fine-grained level towards the discovery of large-scale themes. The data should was read and reviewed to generate an overall sense of its contents. This was quickly followed by line-by-line analysis in which meaningful instances are found and recorded. These instances were considered in the context of the individual's interview(s) and in the context of the overall data set. In this way themes that are interesting and telling at multiple levels can be uncovered. From these general themes we gain insight into additional analysis methods that may be useful for this project, but also additional research directions that may be of interest in the future.

3.5.2 Descriptive Analysis of PALE Log Data

It is straight-forward to perform a descriptive analysis of students' PALE use with information from the PALE log. Some of the basic pieces of information that we can extract in a descriptive analysis include: information about how much time students spent with the system, how many questions they asked whether that varied depending on the type of multimedia support they had access to, and information about attrition and completion. We can also look for facets of the lesson materials that are sources of difficulty for students.

In the past we have estimated the time-on-task for students who worked with the system at home (Nakamura et al., 2010). Now we can compare data for students who used the system at home with students who used the system in classroom and clinical settings to refine our estimation procedures. This quantity may seem superficial, but is actually important. We cannot properly design online activities for students if we do not know how long students will need to complete the activities. PALE logs a timestamp for each action the student makes, but even still evaluating time on task is difficult because of gaps that cannot be understood from the PALE log alone. In 2010 we proposed a saturation-based method for excluding long gaps in student actions to better estimate time-on-task (Nakamura, et al., 2010). The technique is based on the argument that 1) longer time intervals are more likely to be time off task and 2) most of the time intervals are short ones (<1 minute long). One can then set a threshold, say five or seven minutes, above which we say the student is likely off-task and below which they are likely on task. Having completion information for students who used the system in our interview facility has shed light on whether this approach gives a reasonable estimate of time for completion.

We looked at how many questions students are asking the SI tutor as they work through the lesson materials, whether that varies as a function of their PALE treatment, and whether it varies as a function of time. While it is likely unwise to set an "optimum value" for number of questions asked while working with the system, it is likely possible to discern the difference between continuing use of the SI tutor and occasional use of the SI tutor. Our pilot test was plagued by SI tutor under-usage (Nakamura et al., 2010). We made improvements to the system aimed at boosting SI use. This analysis will show whether or not this is the case. Results presented in Chapter 4 do suggest that students query the SI tutor at a level that is higher than typical classroom querying behavior, but lower than one-on-one human tutoring.

An important component of the analysis of the PALE logs was looking for facets of the lesson materials that were difficult for students. Looking at student responses to lesson questions for signs of struggling was the primary means of identifying these stumbling blocks.

Beyond the time and querying information, we can also look at attrition in the PALE log to see whether this is a significant problem. It was problematic in our pilot study (Nakamura et al., 2010). In our current efforts students, at every level, were provided with significant motivation to finish working with the system. Ascertaining whether that motivation was enough to result in significant completion will be of interest.

3.5.3 Using Machine Learning to Group Student Responses

One key question that this project seeks to answer is whether it is possible to simulate the process of tutoring with an interactive video environment that is not based on artificial intelligence. If students do not make use of the SI tutor as a matter of course when working with the lesson materials, then the answer is, No. Based on the pilot study results it seems quite possible that our descriptive analysis will show just that, that many students do not use the SI tutor and for those students the system is not working like a tutor. If that is the case, these data may still be of great use to us, though we would still have further work to do in addressing the question. Graesser and Person outlined a 5-step model of tutoring, which was discussed in Chapter 2, section 2.2.2 (1994). While we will not outline the entire model, it is important to note that our system performs all of the steps except one: the system does not provide brief feedback on the student's attempt to solve a problem or answer a question. This is clearly critical to the system's behavior as a tutoring system. The reasons the system doesn't provide feedback have been discussed previously and are largely due to technical limitations. We cannot use the natural language processing algorithms precisely enough to respond to students' statements the way we respond to their questions. There are two reasons for this, one technical, one practical. The first, technical reason is that the natural language processing routines that match students' queries are not very effective and discriminating based on word order. This is not important for establishing what the sentence is about, but it is very important for establishing whether the sentence is correct. Consider two sentences:

1. Newton's second law tells us that mass equals force times acceleration.
2. Newton's second law tells us that force equals mass times acceleration.

Both of these two sentences are clearly about Newton's second law, mass, force and acceleration. That is easy to determine; the computer can do it. The second statement is, under certain assumptions that are often made in classrooms, correct. The first is not even dimensionally correct. This example clearly illustrates that ascertaining what a sentence is about is easier than ascertaining its veracity. Determining what a sentence is about is largely a matter of looking at the presence or absence of words. Determining what the sentence means requires more subtle interpretive skills or more complex analysis routines. This in a sense makes the natural language processing routines well suited to question and answer exchanges. Even if a student asks a poorly phrased, ill-formed question, the routine will match it based on what it is about, rather than what it means. If the video responses are phrased to address general concepts and then move towards specific instances it is probable that the response will address something related to the student's question, which may enable them to proceed, or to ask a better question, and hopefully get a better answer. If, on the other hand, we want to provide specific feedback on what is right or wrong with a student's response we have moved outside of determining what the statement is about, and moved into determining what it means, the more difficult area.

At this point, however, it is good to remember Chi et al.'s research, conducted with eighth-grade students studying the cardio-pulminory system and tutored by college students, which indicated that tutoring can be quite effective, even when the tutor failed to correctly identify when students did and did not understand the materials (2004). In that study it was found that tutor's overall accuracy in assessing student's beliefs about the circulatory system was about 72%. They conclude that the efficacy of tutoring is more likely based on students' active learning behaviors and, perhaps, an interaction between what the tutor is doing and what the student is doing, rather than as a result of the tutor's focused control and direction of the tutoring process and discourse. Graesser and Person's 5-step model is not in conflict with this finding. It may be possible for the tutor to propel the tutoring session, and provide feedback without necessarily have a deep understanding of the student's understanding, or even the meaning of something they have written. This can often take the form of vague statements like *What did you mean when you wrote this?* or *What do you do after this step?*. Sometimes the tutor may have a clear direction in mind and is attempting to steer the student in that direction; more frequently they may be asking these questions because they do not know what should come next for the student, but they do know that the session must go on, and the student must work towards finishing the problem. The recognition that it may be possible to provide feedback without a detailed understanding of what a student is saying may provide an opportunity for improving the interactivity of our tutoring system.

We have established that it is unreasonable to expect the computer to figure out what a student's statements mean to provide context-specific feedback, but perhaps there is a way to use the student responses to provide less context-specific, but still valuable feedback. The analysis scheme we are proposing to test this hypothesis basically consists of three steps. For each question from our lesson materials, we should have a set of responses from over a hundred students. We can look at these responses and identify similarities in the ideas that the students express. In a sense, we're identifying responses that belong with each other more than they belong with any other responses. We would then like to train a computer model to classify responses to the same question into the groups that we have identified. A utility, called the LightSIDE has been created to do that (Rosé et al., 2007). The software works by first extracting search features from all the responses in a response set. It then looks at the groups and features, looking for statistical correlations that help identify the common features that make responses that we've grouped together similar. Once the model is trained the data is reanalyzed using the model. Higher rates of agreement between the computer's groupings and the human's groupings are obviously better. A model that is successful at this self-check stage (the model is checked against the data that trained it) can then be tested on more data. Again, the figure of merit is inter-rater reliability between the human and computer. If this scheme works, then we now have a mechanism for providing feedback to the students. This process is shown schematically in Figure 3.5.1.

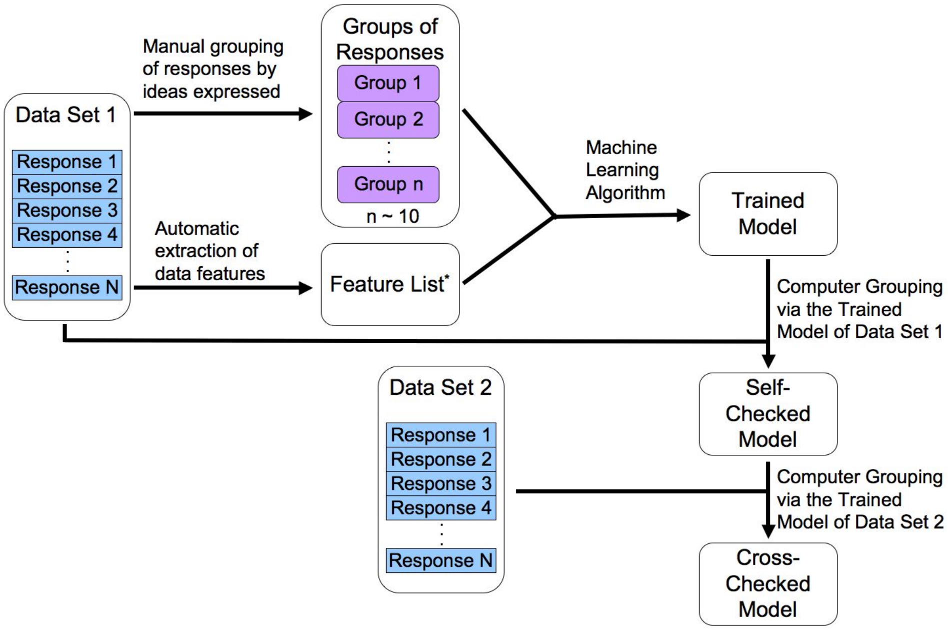


Figure 3.4 Proposed machine learning analysis scheme for student responses to lesson questions

We have already seen some indication that students' responses to conceptual short-answer questions can be broken up into a relatively small number of conceptually distinct groups (Nakamura, Murphy, Juma, Rebello & Zollman, 2009). We have also observed some preliminary success in automatically grouping the responses (Nakamura, Murphy, Christel, Stevens & Zollman 2011). Automatic assessment of short answer responses and its potential for use in providing students with feedback in online instruction have only very recently been realized more broadly in physics education (Butcher & Jordan, 2010; Nakamura, et al, 2011; Jordan, 2012).

In practice using groupings that emerge from a thematic analysis of student responses will likely be more effective than groupings based on our view of what is correct or incorrect, because in that case we are pre-grouping responses that have some latent commonality rather than hoping one will emerge from right and wrong responses. While students may prefer to know if their response is correct or incorrect, it is likely easier and perhaps pedagogically superior to provide feedback based on the ideas expressed and encourage them to extract correctness or incorrectness from physical reality itself.

3.6 Summary

In this chapter a synthetic tutoring system designed to help high school and college students learn Newton's laws and methods of testing the system has been discussed. The system was designed to combine pedagogically sound lessons developed using the learning cycle, with interactive multimedia technology to produce an interactive learning environment. Research designs to test the efficacy of the system while varying the use of the multimedia were discussed. These designs make use of qualitative and quantitative methods including student interviews and the combination of data mining and natural language processing techniques. Research on this type of learning environment has implications for physics education research and physics instruction beyond the immediate scope of this research. The ability to provide socially interactive instruction on-demand and monitor student progress with time resolution is an important capability for researchers and instructors alike. This type of system, when well-understood could play an important role both in the implementation of traditional instruction with online supplements and purely online instruction. We have presented a research protocol that tests the system in three logical settings (clinical interview, classroom, student-chosen). We have also presented a three-part analysis procedure that allowed us to make sense of the data. The four stages are 1) a qualitative analysis of student interview data 2) a descriptive analysis of students' use of the PALE, 3) a machine learning-based analysis scheme that will hopefully yield a method for analyzing large quantities of student responses and allow us to provide feedback. The research protocol and analysis techniques discussed will allow us to obtain a wide range of information about students' use of the system, which will allow us to make inferences about what aspects of the system work best and how improvements should be made. The knowledge we gain from this undertaking could potentially have implications for educators and researchers across a wide range of fields. Understanding how students use our synthetic tutoring system should in turn provide us with a more general understanding of how to construct interactive multi-media based learning environments, an understanding that will be of interest to the Online Learning, Distance Education, and Educational Technologies community. The successful training of computerized models capable of analyzing short-answer questions would mark a significant advance for online homework systems, which can presently only successfully analyze numerical responses and symbolic math. At the same time, this area has not been well-researched, so demonstrating success with this method would certainly imply that further investigation into what we can learn by analyzing students' short answers to conceptual questions is warranted. Determining students' perceptions of the SI tutor and their willingness to work with it also has important implications. While generating the video responses for the SI tutor is quite time-consuming, if it were shown to be highly effective for instruction then, over longer time periods, the start-up work of developing the videos could be offset by reduced or more efficiently directed teaching efforts. These examples clearly illustrate how an understanding of each component of the system can provide educational benefits, just as an understanding of the whole system's functionality can.

Chapter 4 - Understanding Student Experiences with the Pathway Active Learning Environment

In this chapter we present the results of qualitative analysis of semi-structured interviews conducted with 22 students who worked with the PALE system. We also present quantitative results from analyzing the PALE data logs. The students were enrolled in General Physics I, an algebra-based physics course offered at Kansas State University. Eight of the students were male and 14 were female. The students who participated in the study were asked to complete three interviews. Attrition was a minor nuisance and only 59 interviews were completed, with students failing to come to the interview seven times. Ninety minutes were allotted for each interview session. During the session the student spent up to an hour working through one PALE lesson and then spent up to 30 minutes discussing their experiences in the previous hour. The sessions were video and audio recorded. One of the limitations on our design that should be addressed is that we decided, both for ease of implementation and to avoid students feeling "spied on", to video tape their use of the PALE system, but not to actively monitor them (either live or by monitoring the camera signal). Had we monitored them in this way we would have had to disclose that information to them and there was concern that this was neither the most comfortable for the student nor in line with the way the system was ultimately designed to be used. This design choice resulted in missed opportunities to ask students about particular and potentially interesting facets of their use of the system, because the interviewer could not know to ask. Future study may benefit from comparing how students use the system with and without direct observation. Despite this limitation it is believed that a generally clear picture of how students used the system has emerged as well as a clear picture of the way forward in developing this type of system.

4.1 Analysis of Data Logs and Interview Data

This section is divided based on the component of the system that is being discussed. First the SI tutor is discussed, followed by the lesson activities, specifically this means the frame-by-frame measurement because it was the component of the lessons that students struggled with the most. This was clearly indicated both by the PALE data logs and by interview data. The section closes with a discussion of the multimedia support.

4.1.1 Student Perceptions of and Interactions with the SI Tutor

Central to this project is the investigation of how students perceive and interact with our synthetic tutor. Of the 22 students who we interviewed 19 used the system in a configuration that featured an SI tutor (the others used a configuration that just featured the lessons). In collecting interview data with these students it is important to ascertain that the students have had sufficient experience with the system to offer informed comments. That is one of the motivations for conducting interviews over three sessions. Furthermore, in a large majority of the interviews the student either completed the entire lesson within the one hour allotted, or completed most of the lesson activities. The majority of the instances for which most of the activities were not completed were in the first session which was the students first encounter with the system and also contained the additional lesson 0 training activity which required extra time. This indicates that our lessons are at approximately the correct length for students in our target population to work through in an hour, and more importantly that students' comments in the subsequent interviews were based on significant, if not complete experience with the system. As a side note, it also compares well with the method we proposed previously for estimating completion time when direct observation isn't possible (Nakamura et al., 2010). In the following two sections the most important themes from student interviews based on their experiences with the synthetic tutor will be discussed.

4.1.1.1 Social Interaction with the SI-Tutor

Social interaction is critical to effective tutoring and we must look for evidence for or against the idea that students may interact with the SI tutor in a social manner. Moreover we must ultimately look for ways to promote this type of interaction. Evidence for this type of interaction is conflicted. Students frequently discussed the SI tutor as if it was a person ("I asked her about…") however observations based on the video recordings of students using the systems indicate use that is more consistent with it as a video player. Students were observed to rewind, pause and fast-forward the SI tutor or cut it off mid-sentence when it became clear the response was not the one that they were looking for. In a sense this should not be too surprising. Those features are built into the video player and so we must expect students will take advantage of them. The best that we can conclude is that there is opportunity in this type of system for social interaction, but there is no conclusive evidence that that social interaction is occurring.

4.1.1.2 Student Framing of the SI Tutor's Instructional Role

One of the most interesting themes in interviews with students was the idea of using the synthetic tutor "just to make sure" or "to confirm what I knew." Students repeatedly cast their use of the SI tutor in this light. We can see an important and interesting case in the following excerpt, in which a student has claimed to use the SI tutor to confirm what he knew, but what he knew was wrong. At one point in the student's third interview the interviewer asked the student about his use of the SI tutor:

I: *Ok. Ok. But other than that...do you feel like you used the tutor more or less than last time?*

S: *I mean those are two video clips I watched because I was unclear, I watched a couple more while I was doing it just to be clear on what I was doing. I think I knew but I wanted to make sure I was clear.*

I: *mhm*

S: *So I'd say I used it probably the same, but not for the same purposes.*

I: *Ok.*

S: *The times before I'd watch it more because I didn't understand what was happening. This time I watched just to make sure I understood what was happening.*

I: *Ok. Ok. And it reinforced what you were thinking about the physics?*

S: *Yes. Correct.*

This exchange clearly indicates that the student used the SI tutor, or at least believed he used the SI tutor to confirm his understanding. In this same interview the interviewer and the student were discussing the physics of a train crash activity:

I: *So in the first activity we give you these two trains that collide. And we ask you to calculate the force that each one feels right?*

S: *mhm*

I: *Can you tell me a little uh, about what you did there?*

S: *Uh force was just uh, acceleration, no, mass divided by acceleration that's basically how you get how many Newtons of force there are.*

I: *mhm*

S: *And so, that's basically how I found the force of each one.*

I: *Ok. So uh force was mass divided by acceleration?*

S: *Yeah*

This exchange clearly indicates a misunderstanding, or misremembering of Newton's 2nd law. Looking at the log data for this student's user account indicates that he selected the question "What does Newton's 2nd law say?" via the related questions menu. So he was presented with a correct statement of Newton's second law, which would not have confirmed what he already knew. Another related interesting exchange occurs later in this interview:

I: *Ok. Ok. Uhm, so if they felt the same force and had different masses and you're saying that the force is the mass divided by the acceleration right?*

S: *Yes*

I: *Uhm.*

S: *I think. Hopefully.*

I: *Ok, so that's what I'm wondering here. Do you remember Newton's Second Law from last time?*

S: *Newton's Second Law is uhm, about, net force right?*

I: *uh huh.*

S: *All the forces acting on an object, that's the net force, and that's Newton's Second Law.*

I: *And how do we calculate the net force?*

S: *Adding up all the forces being applied, forces this way, that way, gravity, normal force.*

I: *mhm. And what does that equal?*

S: *Well if it's stationary it should equal zero.*

I: *mhm.*

S: *But if it's a non zero net force that means the things accelerating.*

I: *Ok.*

S: *Yeah I think that's right.*

I: *It's accelerating?*

S: *Yes.*

I: *More force means more or less acceleration?*

S: *More acceleration.*

I: *Ok. Uhm...how- when you say that how does that make you feel about the equation force equals mass divided by acceleration?*

S: *That they're interconnected.*

I: *I guess let me rephrase that-*

S: *Yeah.*

I: *If you say force equals mass divided by acceleration, and you make the force bigger, will acceleration- keep the mass fixed- will acceleration get smaller or larger?*

S: *If you make the force bigger?*

I: *mhm*

S: *The mass is the same.*

<long pause>

S: *The acceleration would get smaller, math-wise technically because you could switch the a and the F so it would be m over F equals a. So that way if the force is larger, technically the acceleration should be smaller. I guess yeah that's true...math-wise.*

I: *Which do you have more confidence in? The equation force equals mass divided by acceleration, or your idea that when you apply more force you get more acceleration?*

S: *I mean in my mind, I'm trying to think of a real life example where the more force you apply the faster something would go, but when you look at the math equation it says the opposite of that, so I'm confused now. I don't know.*

I: *Sure. Is it possible you're remembering the equation wrong?*

S: *Yeah. Definitely possible I'm remembering the equation for force wrong. <long pause> Is it mass times acceleration? It might just...it is mass times acceleration, ok.*

I: *mhm*

S: *So, ok. That makes sense now. That makes more sense than mass divided by acceleration.*

This episode highlights a potential pitfall in students' use of our system as well as an example of what a synthetic tutoring system must aspire to. It highlights the former in the sense that we observe a student who is claiming that the SI tutor is just confirming what he knows, but in reality there is a problem with what he knows that the SI tutor could help him with, if he were open to it. The interviewer is successful in helping the student recognize this issue, but the interviewer doesn't do it by providing declarative statements that are in contrast to the students understanding. The interviewer does it by asking the student questions that forces the student to reflect on what he thinks. In this way the interviewer is able to lead the student through a line of reasoning that ultimately allows him to realize his error. This example likely illustrates one of the main reasons one-on-one tutoring is more effective than other means of instruction. Examples like this were not uncommon in the corpus of data, though this is the most striking because of the stark conflict between what the student believed and the accepted physical law and because of the success the interviewer had in guiding the student. Examples where the interviewer tried to do this and failed to help the student reach the correct conclusion were also present in the data. The SI tutor is not currently capable of leading a student through this type of sophisticated and nuanced reasoning chain. One student observed as much indicating that while asking the SI tutor questions might be helpful, the questions that the interviewer was asking her were actually more helpful in getting her to understand things. This is not surprising. In a real tutoring session the tutor doesn't just answer questions she or he also asks the student questions.

4.1.2 Analysis of Student Queries to the SI Tutor

In this section we discuss the analysis of students' use of the SI tutor. Our original intention was to perform this analysis over all data logged with the PALE. A malfunction of the system observed by the conceptual physics students and high school physics students suggest that some queries to the SI tutor (and other interactions with the system) were not logged properly for some students over a brief period of time. It is not known whether the loss of these data was distributed uniformly across treatment groups and we cannot say with certainty how these missing data would skew an attempt to characterize the SI tutor usage. These missing data do not affect our automated analysis (other than having more student responses would likely help the analysis) because we are not trying to make quantitative comparisons across groups, but in characterizing how students use the SI this could make a difference. We are fortunate however, because students who used the system in our interview facility did so while being videotaped and so we have a video record of all their interactions with the system in addition to the PALE logs and we can identify data that is present in the video and missing in the PALE logs and ensure that we have an accurate record of those students' use of the system. Comparison of the PALE logs with the video records indicated that only a small number of students were affected and the missing queries to the SI tutor could be accounted for on video. To ensure as accurate a picture as possible of students' use of the SI, we will discuss only the SI use patterns of students who worked with the system in interviews.

A feature of our system that was implemented as a result of lessons learned from an early pilot test is the ability to either submit queries to the SI tutor via typed natural language or via pre-prepared menu options. Analysis of the PALE data logs shows that three modes of querying the tutor are actually routinely used. These are typed natural language queries, menu-based query selection, and short, keyword-style queries. The first two, of course, are by design. We did not, however intend for students to perform keyword searches and this is potentially interesting as it gives some insight into how students will naturally interact with the system.

Tutor use based on these four methods of interacting is summarized in Tables 4.1 and 4.2. From Table 4.1 we can see that the nineteen students who accessed the tutor in our interview facility submitted 367 queries to the tutor over the course of all the interview sessions. Of these questions 119 were typed into the SI interface, 215 were selected from the quick-start menu and 33 were selected via the related questions menu. With 248 questions submitted via menus, 68% of all questions, the menus were by far the more popular way of interacting with the tutor. Of the 112 typed questions 71 (63%) were natural language questions and the other 48 (37%) were keyword style submissions.

Table 4.2 shows the average number of queries submitted to the SI tutor during each of the three lessons, and in total. Looking at the different tutor-multimedia combinations indicates little difference in the number of queries submitted to the tutor as a function of support media. The case of students whose tutor was supported by static pictures in lesson 2 catches the eye as an unusually large number of queries, and ultimately that treatment group did end up asking the most questions overall. However, given the small numbers in the treatment groups, and the relatively large error bars (quoted as standard error) there is no compelling evidence that there is a real difference between this treatment group and the others. We find that students, on average asked six or seven questions during each tutoring session and that did not change very much for different multimedia support settings. This finding is important because student willingness to engage with the SI tutor is an important condition for successful development of the system. We will return to this result and discuss it in the context of one-on-one human tutoring in section 4.3.

Table 4.1 Querying behavior across all three lessons for all interviewed students who had access to the SI Tutor (N=19).

Question Type	Lesson 1	Lesson 2	Lesson 3	Total
Quickstart Queries	68	79	68	215
Typed Sentences	32	21	18	71
Keyword Searches	10	21	17	48
Related Questions	9	16	8	33
Total	119	137	111	367

Table 4.2 Average number of queries submitted to the SI tutor for the four treatment groups across all three lessons and in total. The indicated error bar is the standard error.

Treatment	Lesson 1	Lesson 2	Lesson 3	Total
Tutor (N=6)	6.3 ± 1.4	6.6 ± 1.9	6.8 ± 2.4	17.5 ± 4.9
Tutor and Pictures (N=6)	6.8 ± 2.0	10.7 ± 2.4	7.2 ± 1.2	23.5 ± 5.5
Tutor and Videos (N=7)	5.7 ± 1.7	6.7 ± 1.9	6.8 ± 1.2	17.3 ± 4.1
All Groups (N=19)	6.3 ± 0.9	8.1 ± 1.2	6.9 ± 0.9	19.3 ± 2.7

Looking at the questions, item-by-item provides some insight into the ideas students asked about while working with the system. While this is more of a qualitative analysis, it is appropriate to discuss it here. The vast majority of typed questions related to ideas like speed, acceleration and force. This is to be expected, given the subject matter. It is interesting to note that, of the 119 queries typed into the interface, only 17 of them (14%) were context-specific questions, that is questions that focused on the objects in the lesson video, as opposed to the physics concepts at play in the video. This is important and interesting because we do not want context-specific queries if we can avoid them. Talking about the physics in the abstract is beneficial in the sense that it limits the number of tutor responses that must be recorded. If we must record a response that discusses Newton's second law in every lesson activity's specific context, that becomes time-consuming. Earlier work on this project suggested that context-specific questions could potentially be a problem (Nakamura, 2010). This work suggests that the problem may not be severe. Another 14 (12%) questions amounted to a request for an equation. This aligns well with students focus on equations during interviews. Eleven questions were related to time, frame rate or measuring time. This is unsurprising because of the difficulty students had with the frame-by-frame measurement. This finding only underscores the importance of refining the frame-by-frame measurement interface in any subsequent research with this system. The last common grouping of questions focused on unit conversion. There were 9 (8%) of these types of questions. While this is not a large fraction of the typed questions, it does suggest that recording responses that deal with unit conversion might be helpful, though one could debate how a real tutor would handle that question. A response of "look it up" would not be completely inappropriate, given the relatively advanced level of these students. Only 4 of the typed questions could not readily be related to the activities and would represent true noise. This very low noise level is very encouraging in terms of further development. While the system has a built in filter designed to reject profanity, which can be adjusted to reject other noisy components as well, it is not 100% effective. Furthermore, working to make it 100% effective may ultimately result in rejecting valid queries. It is far more desirable for students to reject unproductive queries themselves. These data suggest that they do so at a satisfactory rate in the context of our interview setting.

4.1.3 Students' Perceptions and Use of Video Measurement

One of the important features of our system is that students can and must extract information from videos of real physical phenomena and use that information to perform calculations or construct explanations of physical phenomena.

Students' views and experiences with the frame-by-frame measurement were quite varied. It was clear from the interview data that some students framed the lesson activities in a way that was consistent with homework that was to be graded right or wrong. Framing the task in this way may make the idea of extracting information from a video clip confusing at a fundamental level. Extracting information from a video clip provides an extra step in which an error can be made, and these measurements come with inherent uncertainty. Viewed from a perspective that values the infinitely precise numbers from textbook problems this may seem like a negative trade-off. While some students did recognize that the gain comes in the form of making explicit connections with real physical systems, others did not. Moreover there is significant evidence from the PALE data logs that they struggled with it. Observing the video records of the students working with the system shows concrete evidence of how this difficulty manifested itself. Despite the lesson instructions it is evident that some students did not realize that they could move forward and backward through the video clips via two buttons (labeled with arrows) on the bottom of the video player. We had considered this possibility and included an SI tutor response that addressed how to make frame-by-frame measurements using the system. However most (though not all) students did not think to ask the SI that question either by typing it or via menu selection. If we want the students to perceive the SI tutor as an agent that can help them with their technical difficulties, clearly we will have to take great effort to ensure that they are both consciously aware that it can help them with these problems and put them in the habit of using the SI in general. As discussed in section 4.2.1, the possibility of simply forgetting that the SI is an accessible resource is a real possibility.

One of the common manifestations of difficulty with the frame-by-frame measurement involved a moveable position marker at the bottom of the video. This feature, which is common to many online video players, marks the viewers position in the video and allows the viewer to coarsely go forward and backward. In the QuickTime video player that we used in our system, and more generally, this control allows the viewer to navigate through the video on the time-scale of seconds. The frame-by-frame buttons, since our videos were shot at the Internet standard of 30 frames per second, moves the viewer on the time-scale of 1/30th of a second. This position slider cannot provide the same functionality that the frame-by-frame buttons provide. However, students tried to use this slider to make their measurements, particularly in the early activities.

We anticipated difficulty with this facet of the system and integrated a very simple measurement activity in our introductory lesson (lesson 0). In this activity a cart rolls across a ruler and students are asked to measure a distance traveled, the number of frames elapsed during the motion and then convert that number into a time interval. Despite what we believed to be clear instructions, many students did not complete this activity correctly. There is some evidence that students may not have read or internalized the instructions.

An idea commonly expressed in the interviews was discomfort with the inherent uncertainty associated with frame-by-frame measurement. Students framed the lesson activities like homework, not like laboratory experiments. The uncertainty was a source of concern for most students, particularly when connected with the potential impact on a grade for right/wrong answers.

4.1.4 Quantitative Analysis of Student Use of Video Measurement

Looking at the logged data from the undergraduate conceptual physics students and high school physics students who used the PALE in the Fall of 2010 indicated that frame-by-frame video analysis was the primary place where students struggled with the lesson activities. The analysis allows us to develop a quantitative picture student performance on video measurement tasks for a large number of students. Looking at four activities that required frame-by-frame measurement give us a clear picture of the situation. For the simple training activity in which students had to measure the distance that a cart traveled, the number of frames elapsed in traveling that distance and the time interval that those frames equate to we had 133 valid, analyzable responses. Most responses (84 or 63%) were associated with a correct distance, though it is important to note that this measurement only entailed reading a ruler overlaid on the screen. Incorrect responses commonly gave a time interval or said something equivalent to "I don't know." In terms of the frames elapsed, only 38 out of 133 or 29% were within a reasonable margin of error. Of these responses 29 calculated a correct time interval. This is 74% of those 38 and 22% of the total response set. In the next activity, the first in the real lessons, students had to use frame-by-frame analysis to measure the speed of a ball at the beginning and end of a track. On this activity we had 156 valid, analyzable responses. Of these responses 13.5% did not actually submit numeric answers for the speeds. Forty-three responses indicated a number within 10% of the expected values, which is about 27% of the total response set. One response indicated a failure of the video to work properly. The next activity that required frame-by-frame analysis was the third activity in lesson 1, in which students were asked to measure the speed of a moving car and a crash test dummy before and after a crash. This activity included 136 analyzable responses. Most (115 or 83%) of these responses correctly indicated that the car's final speed was zero. Only 36 (26% of the 136 total) responses contained a number for the car's initial speed that was within a reasonable margin of error. This measurement was slightly more difficult due to quality of the video, so a generous 30% margin of error was considered. Surprisingly 11 responses indicated that the car's initial speed prior to the crash was zero. Measuring the dummy's speed was still harder, but 31 responses contained a numerically reasonable value. The picture is clear: with the current set-up we can expect 20-30% success rate for fairly simple measurements. For more complicated measurements it becomes more challenging to ascertain whether the student does not understand the measurement process or does not understand the broader context in which the measurement is being done. For example in the second lesson students were asked to apply Newton's 2nd law in the context of a softball being hit. In this activity the students were to measure the velocity of the incident softball (a distance and time measurement), measure the velocity of the receding softball (a distance and time measurement), recognize that they can use the change in velocity and the interaction time with the bat (which they can estimate from the video) to calculate the acceleration and then use Newton's 2nd law to calculate the net force, which is due to the bat. On the one hand this seems like a rather complex series of computations, but at the same time this is a very standard textbook problem in introductory physics, so the community has, in a sense, decided that this is an appropriate problem for introductory physics students. And certainly some can do it, but doing it in the context of video analysis may add an extra complexity that they are not ready for. With 118 unique responses in the data log we observed only 3 reasonable values for the force exerted by the bat on the ball. The picture is slightly better for measurement of the ball's incident and receding velocities. Here 19 responses contained reasonable values for the velocities (16% of total). What is troubling is that 26 responses were not numeric suggesting that these students could not, or would not perform the measurement requested. This calls the remaining responses into question at some level. How many are just guesses and how many are bad measurements? It is difficult, if not impossible to tell. Incentive then becomes a relevant concern at this point. Since students reaped rewards for completing the lessons and did not suffer consequences for getting the answers wrong it is difficult to tell whether they mostly made their best effort to perform the measurements. While this is a potentially confounding factor, it is clear that a success rate this low for frame-by-frame measurement is not good enough for the current implementation to be considered successful, even if some students' effort is lacking. It also seems probable that the success rate is too high to suggest that none of the students were taking the measurement seriously. We are left to conclude that the system must be modified to improve this facet of the lessons. This picture of students' difficulties with video analysis is combined with our understanding from discussions with students in interviews to create a more holistic picture of the situation in section 4.4.2. Methods to improve the system and make video analysis a more viable component of the system are also discussed.

4.1.4 Student Perceptions and use of Multimedia Support

Establishing a clear set of criteria for determining the efficacy of the multimedia that supports the SI tutors' responses is quite challenging. We can make some inferences, however based on our observations of students' use of the system as well as their comments about their use of the support media and its perceived value to them. While this approach will clearly not satisfactorily answer the question of how useful this feature is in instruction it provides initial insights that can help us construct future research designs that can answer the question in a more satisfactory manner. Most students who had access to images or video clips responded positively to them. A small minority identified them as distracting. While students mostly did respond positively they could not always remember what they had seen or if they had seen anything at all. Observations of student use of the system on video showed sporadic examples of clear distraction, though due to the positioning of the camera (necessarily behind the user's head) it was impossible to tell with great certainty whether the student was focusing on the support media. Students were observed to pause, fast-forward and work through the support videos, though others reviewed support videos a second time, suggesting that they were getting something out of it. We do not currently have a good metric for identifying whether one of these modes of multimedia support genuinely promotes learning. One quantitative metric that can be applied is a comparison of querying behavior based on support media. If students were more or less likely to query the tutor based on the multimedia support they assigned we would have evidence that the type of multimedia affected overall interactions with the system. However, from Table 4.2 we can see that the querying behavior for interviewed students is independent of multimedia support to within error bars. Eye-tracking has emerged as an increasingly interesting means of connecting the things students are looking at with learning (Madsen, Larson, Loschky & Rebello, 2012; Smith, Mestre, & Ross, 2010). This approach should be considered for future study. At the very least it would provide more information about how much students attend to the multimedia support. This information comes at a cost in terms of an additional level of complexity of the set-up, and careful consideration is required to ascertain the potential benefits.

4.2 Discussion of Results

We have observed that algebra-based students asked an average of 6.3 questions per hour while working with our system in our interview facility. Graesser and Person investigated question asking in tutoring and observed an average of 26.5 questions per hour in one-on-one human tutoring (1994). They also state that classroom questioning occurs more along the lines of 0.11 questions per hour (Graesser & Person, 1994). This would indicate that our system promotes questioning far beyond the classroom level, but not yet at the level of one-on-one human tutoring. This should not be surprising and should be viewed as an positive indicator of students willingness to interact with the synthetic tutor and use it as a means of getting information. Generally students did ask questions of the SI tutor as a means of getting information. Increasing the interactivity of the tutor via methods like those discussed in Chapter 5 might further promote querying. We looked for evidence of students interacting with the tutor as a social actor, and many do refer to it like a person, but there is significant evidence from video observation is that they interact with it as a video player. A valid question that should be asked is whether this interface should allow students to pause, rewind and fast-forward the video responses. The extent to which one can perform conversational analogs that with a human tutor (by interrupting, asking for repetition, or omission) is governed by social norms and rapport. It likely varies considerably in real human tutoring interactions. One can make the argument that this is a decidedly positive feature not a negative one (and several students did comment on that), but at the same time it is important to consider how this set of features affects the system's capacity to behave like a human tutor. If we want students to interact with it like it is human we will have to build social norms into that interaction. It cannot be a tutor when we discuss it and a video player when we use it. Specific investigations of how control over the tutor's speech affects students interaction with and use of the tutor may be warranted. Also of interest is the question of how students are using the tutor, as discussed in the next paragraph.

When addressing the finding that multiple students considered the tutor to be useful to confirm their existing understanding (an understanding which may very well be incorrect), it is natural to ask whether it is possible to modify the way the system functions to promote the tutor as an agent that encourages students to confront their existing knowledge and determine whether it is, in fact, consistent with observations. When we consider either Chi's research on the efficacy of tutoring or Van Lehn's model of the tutoring interaction, clearly pushing students towards making this comparison is likely an important facet of the tutoring method. In Chapter 5 we investigate a method for providing students with automated feedback based on their responses to lesson questions. That method may provide a means of prompting students to reflect on their understanding and allow the system to function more like a real tutor. In the absence of that method it is worth careful consideration to determine if the lessons and the environment as a whole could be restructured to permit more questions being asked of the student by the tutor, even if those questions are not based directly on students' prior actions.

The video analysis component of the system is another area where refinement is likely required. While the picture from the data logs seems suboptimal, observing students' use of the system on video as well as interview data indicated that students can discover how to use the frame-by-frame buttons along the way, even if they have neglected the instructions, and they can understand the measurement process and how frame-rate is connected to time intervals. However they are not used to thinking in this way. Many of the students whom we interviewed framed these activities as a kind of online homework and homework has precise givens that are processed to yield precise unknowns. By attempting to make activities that better connect to real life we have drifted away from a comforting idealization and introduced an additional complexity. At the same time, getting quantitative information from real physical systems is at the very heart of physics and should therefore be an important component of this system. It is difficult to know with certainty whether students struggled with the video measurement component of the lesson activities because they did not understand the method or because they were unwilling to perform the task. It almost certainly varies from student to student. Reducing the barriers to successful completion of the video measurement via instructions which are more demonstrative and less explanatory or via an interface which is more intuitive could both improve the situation for students who did not understand the method or for students who felt it was not worth the effort.. In terms of improving the interface, two possible solutions are clear at this point. We have recently experimented with putting a visible time stamp on each frame of video so that students can more easily make time measurements without counting frames and converting to time intervals. This somewhat effort-intensive, but removes the additional level of complexity introduced by requiring students to convert from frame number to time interval. Another possibility is to explore adjustments to the video player to make it more like other established utilities for video analysis, such as Tracker (Brown & Cox, 2009). There is evidence that high school students can do frame-by-frame measurement without that type of interface (Brungardt & Zollman, 1995). At the same time we do not have the advantage of being able to deliver instructions via in-person communication. Our instruction and interface must be sufficiently transparent and simple that students can navigate the system essentially unaided.

The video player that we used was chosen based on the following criteria: it supported frame-by-frame navigation and did not require extensive development by our development team at Carnegie Mellon University. For technical reasons related to improving data download speed virtually all streaming video, which is increasingly becoming the standard on the Internet, do not support frame-by-frame navigation. The only video player that met these criteria was QuickTime. It is interesting to note that the relative prevalence of streaming video may be connected to students' difficulty with this facet of our system. With the advent of YouTube it is impossible to believe that the participants in our study were not familiar with video on the Internet and several said explicitly that they were frequent users of web video, but that is increasingly streaming video, which is different from the QuickTime video player that we used for that portion of the system. It is clear in retrospect that these students likely did not think of frame-by-frame navigation as something commonly done in Internet video. It definitely possible to develop an interface that lends itself to doing frame-by-frame video analysis in a more intuitive way that can be built into our system. This would probably look like a scaled-down version of the Tracker software that is commonly used for video analysis (and which cannot currently be embedded in web applications) (Brown & Cox, 2009). While our development team originally hoped to avoid this type of endeavor we can clearly see from our observations of students that it is almost certainly necessary for the successful development of this type of system.

At the same time effort should be put into helping students understand and accept the uncertainty associated with video analysis and allay any concerns about getting the wrong answer. Because assessing numerical results is comparatively easy, it would likely be of considerable benefit to the system to modify the system so that questions that admit numerical answers result in automatic feedback when the numbers are outside of a certain range. If this feedback were to be provided by the SI tutor, it is possible that this could boost student use of the tutor. Evidence that students may not carefully read instructions suggests that another means of conveying this information is necessary. Using the SI to provide instructions is one option, explicitly demonstrating the techniques involved in making the measurements, and what constitutes acceptable uncertainty is another.

We do not see evidence that the particular multimedia that supports the SI tutor's responses affects students querying behavior for the students that we interviewed. While video support has a higher degree of interactive capability associated with it, as compared to static pictures, in the sense that students can review and fast-forward through the video to focus on the information that they perceive to be useful video are also more time-consuming to create. Students were observed to interact with video support in this way, but it is difficult to conclude that this interaction supports their learning. Developing metrics that can conclusively establish student engagement with the multimedia support and allow better comparisons between different multimedia modes is a priority for future research efforts. Even ascertaining whether students are looking at the multimedia is difficult in observing video recordings of student use. Eye-tracking technology may provide us with one quantitative measure of which kind(s) of multimedia support are most engaging for students. Early testing suggested that a pre-test/post-test design was not an effective means of comparing multimedia for this type of short intervention (Nakamura, 2010). However, it may be useful to extend the number of lessons, and thus the duration of the treatment, to make the pre-test/post-test design a more viable means of comparing the different modes of multimedia support.

4.3 Summary & Conclusions

Analysis of student interview data and PALE data logs have provided a direction forward in the development of a video-based interactive tutoring system. We see evidence that students interact with the tutor by querying it that is greater than in a typical classroom environment but not at a level consistent with one-on-one tutoring. Students' queries to the SI tutor were largely productive with little noise or inappropriate content. Students were generally satisfied with the tutor's responses to their questions but is some evidence from interviews that indicates that they may be easily turned off by responses from the SI that don't match their queries and that they aware of its potential limitations. This is one possible explanation for the relatively high rate of menu use as compared to typed questions, though some students indicated uncertainty about what to ask. Students have expressed a desire for feedback on their work and concern about uncertainty and correctness in the video analysis central to many of our lesson activities. Using the SI tutor to provide feedback would be desirable. One potential method for doing so is investigated in Chapter 5. Using the SI tutor to question students about their understanding is also desirable. The method explored in Chapter 5 may also assist with developing this capability.

Chapter 5 - Automated Analysis of Short Text Responses

5.1 Automated Analysis of the Short Responses to Lesson Questions

In this chapter we present the results of using a learning algorithm to develop machine-learning models capable of automatically assessing student responses to short-answer questions contained within the PALE lessons. The responses analyzed were collected from several groups of students. The algebra-based physics students whose interview data was discussed in the previous chapter are included. Responses from the group of 30 algebra-based physics students who used the system in the Summer of 2011 are also included as are responses from 41 high school students who used the system in the Fall of 2010 at the request of their teachers. Ninety-nine students enrolled in a conceptual physics course targeted at elementary education majors used the system in the Fall of 2010 as well. While these groups of students are different in many ways, the physics that they study and the ways in which they study it substantially overlaps with the target group. It was determined that combining sets of responses from these different groups of students was the best possible way to obtain a maximum number or responses, which is typically important for success using data mining analysis schemes. Attrition, or mortality was an issue in this research; while the above numbers suggest that approximately 200 students worked with the system we could not ensure that all of these students completed all of the materials and in practice the number of analyzable responses is typically much less than the total number of students who worked with the system. This danger is inherent in doing research with online instruction, where it is often difficult if not impossible to control student completion rates. The analysis procedure is discussed in Chapter 3 section 3.5.3 and depicted in Figure 3.4. Here we will briefly review our motivation for this analysis and discuss the details of our implementation. We will also present the results of the analysis and a discussion of those results. The chapter ends with conclusions that we can draw from this analysis and its discussion.

5.1.2 Automated Assessment Analysis Procedure

The purpose of this analysis is to ascertain whether using machine learning to classify students' responses to our short-answer lesson questions is feasible. Being able to do so advances the project in two important ways. The first is that being able to automatically classify a student's answer to a question opens up the possibility of automatically providing feedback to the student based on that answer. The ability to provide feedback is one of the five roles of a tutor identified by Graesser and Person (1994), and it is a role in which our system performs weakly. Our system does provide students with feedback, but that feedback is not immediate and it is not tuned to each student's work. Being able to automatically interpret a student's answer is clearly the first step to providing this customized feedback. The second benefit of this approach is that it may provide us with a more structured framework for understanding students' progress through the system. Grouping student responses provide discrete labels that may reflect our insight into how students are responding to questions within a given activity. Looking at how those labels for individual students, or for larger groups of students who have completed multiple activities, could quite possibly provide more insight into how students are progressing through the system.

The responses that we analyzed came from a variety of students in a variety of educational environments, as discussed in Chapter 3. Students who used the system were enrolled in one of several high school classes, a conceptual physics class at Kansas State University or one of two algebra-based physics classes. While high school students, conceptual physics students and algebra-based physics students are different in many ways there is also evidence for similarities. Review of the response data indicated significant commonalities in the ideas expressed by the three groups of students.

To analyze the data we begin by looking at the data logs to ascertain which questions from which activities are most appropriate for this sort of analysis. It should be clear that not all questions will produce response sets that break down naturally into coherent groupings. While most of our lesson activities are designed to address concepts as well as calculation some focus more on one of these or the other. It should be clear that activities that focus more on concepts are more appropriate for this type of analysis than activities that focus more on calculations. Activities that focus on calculation are already more-easily assessed automatically by virtue of the answers being numerical. Nine questions were selected that seemed to present conceptually distinct groupings which might predict success in this type of analysis.

Once responses sets were identified as being appropriate for this type of analysis, some pre-processing of the data was necessary. Identical duplicate responses (which can show up if a student submits answers twice in one of several ways) were removed. Inappropriate or irrelevant responses were also removed. The repeated responses and irrelevant responses were infrequent, typically less than a few percent of the total response set. The responses were spell-checked with the commercial spellchecker included in Microsoft Office. A person determined at each instance whether a correction was to be made. If no clear intended word could be identified the word was left as the student typed it. Common modern abbreviations such as idk (I don't know) and b/c (because) were exchanged with their formal equivalents. While most punctuation was left as the student typed it, commas were searched and removed. This was because of a potential formatting issue with comma separated data files, which our analysis program makes heavy use of.

The response sets were again read and re-read several times to get a better feeling for the ideas that students expressed in answering the questions. The responses to an individual question were grouped together based on commonalities in the ideas that were expressed. These groupings should not be viewed as unique. They should also not be viewed as correct or incorrect. As will be discussed in section 5.1.3.2, several response sets yielded multiple grouping schemes that are reasonable to adopt. It will become clear that different groupings may provide different benefits but also different disadvantages. Our goal is not to exhaustively find all the ways of coding these responses, but instead to assess how reasonable the analysis approach is and whether it provides useful insight into how students use the system.

Once the responses had been manually coded they were ready for analysis via the LightSIDE utility. Each response set was analyzed twice using LightSIDE. In the first analysis the entire response set was used to train a model that could be used to code additional data. In the second analysis the responses were ordered using a random number generator and then divided in half. Each half was used to train a model, and each model was used to code the other half. Each model that is trained yields information about its performance in the form of a self-check. Each set of data that is coded via a model provides cross-check information about the performance of the model. The reason for training a model using the entire response set in addition to the second method is that training on more data generally produces better models. It is useful to compare the self-check information from the model trained on all the data with the information we obtain from the second approach. If we observe that the performance of the model trained on the full data set is significantly higher than the performance we observe when coding half of the data with a model trained on the other half we can conclude that our data set is likely too small to divide in half and still obtain useful results. While this result would be disappointing in some sense, it is useful because it sets a lower limit on how many responses one needs to do this type of analysis. Conversely, if we observe general agreement between the results of the self-check on the full data set and the cross-check of the half data sets we have evidence that the self-check results predict the cross-check results, which would make logical sense. Seeing that result is beneficial for future implementations because it provides us with information about how much cross-checking of a model is necessary. It may not be necessary to build the model, self-check it and then cross-check it before deciding whether to collect additional responses to cross-check it again. The self-check alone may provide sufficient information to know whether the model works well enough to warrant continued use.

Training the model begins by extracting a feature list from the set of coded responses. A feature is essentially any searchable entity in the data set. These can include single words (unigrams), punctuation, groups of words (bigrams, trigrams, etc.), line length as well as other possibilities. We have observed the best results with feature lists consisting of single words (unigrams), line length and punctuation. Feature extractors commonly allow the user to remove so-called stopwords from the feature list. These are common words that are not likely to be a predictor of any of the classifications. We have generally used this function in LightSIDE to reduce the size of the feature list, which speeds up the model training process. We have not observed a significant improvement or degradation of the trained models when the stopwords were removed. The feature extractor also extracts information about how the features are distributed across the responses set. This information is critical to the analysis because it allows us to understand the features as predictors of a response belonging to a given group. In extracting this information features can be treated as binary or continuously distributed. If the features are treated as binary, then only their presence or absence in each response matters. If they are treated as continuous, then the frequency with which the features shows up in each response matters. Given the shortness of the responses in our data set it is more common for a feature to show up in a response or not. Therefore treating the features as continuously distributed in our data set does not result in significantly different results as compared to treating them as binary. However, features were treated as binary for the analysis presented in this dissertation.

The extracted feature list is used to train the model via one of a number of learning algorithms. We have experimented using Naïve Bayes, classifier which has been demonstrated to be useful for text classification (Kim, Han, Rim & Myaeng, 2006). We have also experimented with an implementation of Support Vector Machines (SVM) called Sequential Minimal Optimization (SMO) (Platt, 1998). These are both included in LightSIDE and are two of the most commonly used classifiers for this type of work. Our initial work in this direction used the Naïve Bayes classifier because it ran faster on our machines, but both SMO and Naïve Bayes gave very similar results (Nakamura, 2011). It is not surprising that the Naïve Bayes implementation ran faster, as SVM is significantly more complicated an algorithm (Kim et al., 2006). For technical reasons having to do with how the newest version of LightSIDE's Naïve Bayes classifier treats binary features we performed all the analysis presented in this dissertation using SMO. This allowed us to continue to treat our features as binary and use what is likely the more common, if slightly slower, tool for this type of analysis.

5.1.2 Success Criteria for the Analysis Scheme

To assess how well this method performs in classifying responses via the trained models it is useful to look at the results from several perspectives. The first and last judgment on the efficacy of the method comes in the form of inter-rater agreement between the human who generated the coding scheme and the computer that was trained on data coded with that scheme. The most basic measurement of this agreement is the percent of responses that were coded the same way by a human rater and the computer model. Scoring a high percent agreement is certainly a goal of this approach, but there is at least one drawback to this metric. Percent agreement does not take into account the possibility of random chance agreement. Another statistic that takes this possibility into account is Cohen's kappa (Cohen, 1960). Cohen's kappa is the ratio of the difference between the observed match rate and the expected random match rate divided by one minus the expected random match rate. When two raters match at a rate that is consistent with random matching, kappa is zero. When two raters match perfectly, Cohen's kappa equals one. A scheme for interpreting Cohen's Kappa is commonly taken from Landis & Koch's 1977 work on inter-rater agreement. While the author's themselves assert that their classification scheme is somewhat arbitrary and useful for providing a more intuitive feel for certain examples, it has none-the-less become somewhat of an accepted standard for inter-rater agreement (Krippendorf 1980; Nehm & Haertig, 2010). The framework is shown in Table 5.1

Table 5.1 Benchmarking demarcations for Cohen's Kappa values

Kappa Value	Agreement
<0.00	Poor
0.00-0.20	Slight
0.21-0.40	Fair
0.41-0.60	Moderate
0.61-0.80	Substantial
0.81-1.00	Near Perfect

Beyond inter-rater agreement, it is useful to look more closely at how the trained model assigns responses to different groups and what happens when it disagrees with the human coder. A simple way to do this is to look at the ways in which the computer model is misclassifying responses, to ascertain whether a pattern emerges.

5.1.3 Results of Automated Analysis Procedure

In this section the results of the automated assessment procedure are presented. These results are divided into two sections. In the first section the agreement between the computer model and the human coder is presented. In the second section observations of the ways in which the model fails are presented.

5.1.3.1 Agreement Between the Computer Model and Human

Table 5.2 summarizes the principle results of our analysis. It shows the overall match rate that was generated by doing a self-check on models trained on all the analyzable responses to a given question as well as the match rate that was generated by splitting the response set in half, generating two models and doing two cross-checks. Cohen's Kappa is also reported for both match rates. Nine questions were analyzed with LightSIDE as described above.

The first thing that is noteworthy about these results is that the match rate obtained from the self-check of a model built with all the responses is a good predictor of the match rate generated via cross-check using two half-size data sets. We would expect that the cross-check match rates would be reduced with respect to a self-check match rate for two reasons: the model is trained on less data, and it is trained on different data. Therefore a high cross-check match rate should be looked upon as meeting a higher standard of validation for the coding scheme as applied to the data set than a high self-check match rate. Across the board we do see a reduction in the match rate moving from self-check to cross-check (or in two instances no change), but higher cross-check match rates generally go along with higher self-check match rates.

Table 5.2 Match rate for machine learning analysis method

Lesson Question	Groups	Total Responses	Matched (self-check)	Kappa	Matched (cross-check)	Kappa
Ball & Track Q5	6	161	116 (72%)	0.6235	112 (70%)	0.5791
Car & Coffee Cup Q2	14	154	75 (49%)	0.4205	55 (36%)	0.3614
Crash Test Dummy Q5	6	161	136 (84%)	0.6535	135 (84%)	0.6743
Coin & Cylinder Q1	9	150	109 (73%)	0.6714	92 (61%)	0.5328
Beaker & Coin Q2	9	142	76 (54%)	0.4654	54 (38%)	0.3024
Hammer & Feather Q1	5	158	140 (89%)	0.7947	134 (85%)	0.7128
Train Crash Q2	15	105	52 (49%)	0.3888	38 (36%)	0.2282
Ice Skater Q4	8	110	77 (70%)	0.6298	77 (70%)	0.5653
Live & Dead Ball Q2	9	89	49 (55%)	0.4396	44 (49%)	0.3321

The second interesting feature that emerges from the table is that the questions break down in to two distinct groups. Five questions could be classified via groupings that resulted in self-check kappa values of at least 0.6200 and match rates of at least 70%. Four questions resulted in self-check kappa values of less than 0.4800 and match rates of less than 60%. The former group would be considered substantial agreement by the benchmarking standards we've adopted, while the latter group is only moderate.

The third useful observation that we can make from the table is that questions for which it was possible to group the responses into fewer groups generally presented higher match rates and correspondingly higher kappa statistics. Questions for which the responses could only be grouped into larger numbers of groups resulted in poorer matching rates. This is quite reasonable. In a very real sense the number of groupings that emerge from a response set is a measure of the conceptual coherence in the response set. We can imagine a continuum on one end of which is the situation in which all students say the same thing and there is only one group, and on the other end is the case in which every student says something different and we have the same number of groups as we have responses. We have established that the groupings are not unique, and it is possible that re-analyzing the response sets could reduce the number of groups in some of these questions. This is somewhat unlikely because the present analysis was conducted with parsimony as a goal that was secondary only to ensuring that responses within groups really reflected the commonality identified. Trying to scale down the number of groups also reveals an inherent balance that must be struck in this type of analysis: reducing the number of groups will generally increase the likelihood of grouping responses that really do not belong together. So while we cannot make the strong claim that these groupings represent a true optimum, we can make the weaker claim that working to reduce the number of groups will likely come at a cost in terms of reducing our ability to describe conceptually distinct ideas expressed by students. The trade-off between creating a framework capable of capturing nuances of different responses and creating a conceptually simple framework can likely only be balanced on a question-by-question basis as more experience with this approach is established.

5.1.3.2 Further Discussion of Select Questions

The reason that the automated assessment technique did not work very well for the four questions in Table 5.2 that yielded match rates below 70% almost certainly centers on the large number of groups used to analyze the responses and the lack of conceptual focus in students' responses. These are, in fact, two sides of the same issue. A more important question for our goal of understanding how to use this method to provide students with feedback as they work with our system is how does the approach mismatch when applied to the other five questions. In the following five sections the details of the mismatches for each of the five most successful questions will be discussed. The discussion will be in the context of the computer model trained on all of the data. Therefore the matching and mismatching will be discussed in the context of the self-check of the model, not a cross-check. As we have seen, there is evidence that the self-check provides insight into the behavior of the cross-check, and this is particularly true for these better matching questions.

5.1.3.2.1 Lesson 1 Exploration Activity 1 Question 5

In this activity students measured the speed of a ball at the beginning and end of a flat track to observe that it did not change significantly. The students were then asked to make inferences of the speed in the middle of the track and at the end of a track that was twice as long. The last question, which focused on the speed of the ball at the end of the longer track proved to be a good question for the automated assessment scheme. In Table 5.3 the groups that were used to analyze the data are listed along with the number of responses in each group, the number and percent matched by the computer model, and the number and percent mismatched. So, for example 58 students said that the ball would be slower at the end of a longer track, and the computer model successfully matched 48 of those responses (83%). Viewing the data from this angle suggests a very positive facet of the analysis. If a student said the ball would be slower (which is not consistent with the video clip), the computer would correctly match that 83% of the time and we could provide the student with a video clip prompting them to reconsider the answer. It is useful, however to consider the inverse of match rate. The interesting thing about looking from that perspective is that we can observe that the computer model tagged 68 of the responses as belonging to this group. This is twenty additional responses erroneously tagged as belonging to that group, 29% of all responses so tagged. If this rate persisted with additional training and data collection, we could count on nearly thirty percent of all students who received the prompt to get it in error. Investigating the actual matching of the responses showed, not surprisingly, that seven of mismatched responses were coded as "Slower but not much" by the human. This is a rather nuanced distinction and whether the computer can make this distinction is one of the things we seek to clarify. We should not be terribly surprised if it cannot. A more disturbing observation is that the next most common mismatch to that group was from responses that had been tagged as the ball having greater speed. These six mismatches explain the bulk of the models failure to correctly classify responses that indicated the ball would speed up. Overall the model did a better job with responses that indicated the ball would maintain the same speed with only slightly more than ten percent contamination, with that coming fairly uniformly from all the other groups. We can therefore say that in this case the model does a better job of identifying responses that are more in line with what is observed in the video clip. This suggests that in this case it may be better to ascertain what type of feedback a student gets based on whether the response was tagged as being consistent with the ball moving with the same velocity.

An interesting observation on this question relates to the effect of being able to reduce the number of groups used to analyze the responses. Prior to using the coding scheme shown in Table 5.3 a finer grained coding scheme was attempted. It was observed that a significant number of students expressed the idea that the speed would be half as great, or twice as great. It was also possible to resolve between students who said that the ball would be slower and those that said it would be *much* slower. This coding scheme resulted in nine groups instead of six. It also resulted in an overall self-check match rate of 63%. Recognizing that some of these groups could be collapsed together without degrading the quality of the coding scheme allowed for significant improvement of the model-human inter-rater agreement.

Table 5.3 Match/mismatch details for lesson 1 exploration activity 1 question 5

Group	Number	Matched	Percent	Tagged	Mismatch	Percent
Slower	58	48	83%	68	20	29%
The same	43	39	91%	44	5	11%
Slower but not much	20	11	55%	16	5	31%
Greater	18	8	44%	14	6	43%
Number focus	15	9	60%	17	8	47%
Same or Slower	7	1	14%	2	1	50%
Total:	**161**	**116**	**72%**	-	-	-

5.1.3.2.2 Lesson 1 Exploration Activity 3 Question 5

This activity focused on the motion of an unrestrained dummy in a crash test. In the video clip students observed a car and dummy moving with the same speed prior to a crash and then observed that once the car came to a rest the dummy maintained the same speed. Ideally they measured the speed. In question 5 they were asked why the dummy's motion ultimately stopped. In this case the vast majority of responses focused not on physics concepts, but on describing what happened in the video clip. Here the computer model matches a high rate (98%) with relatively little contamination (8%) from other groups. While at first this may seem somewhat trivial, since the students' responses are not indicative of a common misconception or mistake, there is in fact potential here. If a tutor asked this type of question and a student simply described the situation, the tutor would almost certainly prompt the student to go further and connect the physics concepts to their description. If a model can correctly classify responses that consist of this type of description instead of explanation then the SI tutor can also provide this prompt. Unfortunately the model was less effective when it came to responses that discussed the dummy feeling a force (which is the accepted explanation). It correctly matched 91% of the responses that used this explanation, but did so with significant contamination, mostly from responses that discussed force, but not in the context of a force on the dummy.

Table 5.4 Match/mismatch details for lesson 1 exploration activity 3 question 5

Group	Responses	Matched	Percent	Tagged	Mismatched	Percent
Literal physical description	104	102	98%	111	9	8%
Force exerted on dummy	34	31	91%	45	14	31%
Other force-related ideas	16	3	19%	6	3	50%
Newton's 3rd law ideas	4	0	0%	2	2	100%
Momentum transfer ideas	2	0	0%	0	0	-
Acceleration ideas	1	0	0%	1	1	100%
Total	**161**	**136**	**84%**	-	-	-

5.1.3.2.3 Lesson 1 Application Activity 1 Question 1

In this activity students were asked to apply Newton's first law to the task of obtaining a coin stuck lightly in the bottom of a graduated cylinder. This is very similar to the dummy crash test activity in the sense that the correct approach is to crash the tube into the surface of the table. The coin, which is in motion prior to the crash, will maintain its motion and come free. The groupings, match rates and mismatch rates are shown in Table 5.5. It is interesting to note that many responses indicated that a force was required to obtain the coin. Many students who submitted responses to this question that did not explicitly mention force (solutions suggesting hitting the bottom of the cylinder, or hitting the cylinder on the table) clarified in later questions that they believed that getting the coin hinged on applying sufficient force to the coin, rather than exploiting Newton's first law. The relatively high match rate across the top half of the groups would suggest that it might be possible to preemptively challenge that idea. At the same time the relatively high amount of contamination, which could not be attributed to a small number of clear sources, but instead was observed to be fairly uniformly distributed, must be overcome

Table 5.5 Match/Mismatch details for lesson 1 application activity 1 question 1

Groups	Number	Matched	Percent	Tagged	Mismatch	Percent
Invert the cylinder	31	31	100%	43	12	28%
Hit the bottom of the cylinder	30	21	70%	31	10	32%
Hit the cylinder on the table	26	18	69%	23	5	22%
Apply force to the coin/cylinder	24	22	92%	29	7	24%
Use gravity to get the coin	10	10	100%	10	0	0%
Other ideas	10	1	10%	5	4	80%
Proposes multiple methods	10	3	30%	6	3	50%
Inertia ideas	6	2	33%	2	0	0%
Shake the cylinder	3	1	33%	1	0	0%
Total:	**150**	**109**	**73%**	-	-	-

5.1.3.2.4 Lesson 2 Application Activity 1 Question 1

This activity centered on the dropping of a hammer and feather on the moon. In this question students were asked to predict what would happen when the astronaut dropped the hammer and feather (simultaneously). In the remainder of the activity they also performed various calculations designed to encourage them to apply Newton's second law. The groupings, match rates and mismatch rates are shown in Table 5.6. The analysis of these data represents the best case for using this approach for providing students with feedback. The first very unusual thing can be observed from the data is that a significant number of students did not know that there is gravity on the moon. The fact that a significant number of students thought the hammer would fall faster is not surprising. The fact that the computer model correctly matched responses to those three groups with high accuracy and little contamination is very encouraging that we could provide a prompt asking students why they thought the hammer would fall faster or whether they were aware that there was gravity on the moon. This activity provides the best evidence that it is possible to provide feedback to students by the method investigated. We see clear conceptually distinct groupings of responses, and we see that the computer model does a reasonable job of classifying responses based on those groupings. We can envision a real tutor asking a student whether they know that there is air resistance on the moon, or why they believe that the hammer should fall faster. For the three largest groups we would expect to succeed with this type of intervention more than 90% of the time, which would be a good starting point for further development

Table 5.6 Match/Mismatch details for lesson 2 application activity 1 question 1

Groups	Number	Matched	Percent	Tagged	Mismatched	Percent
The fall the same	98	93	95%	101	8	8%
They float	27	23	85%	25	2	8%
Hammer falls faster	22	20	91%	22	2	9%
They fall (Phys. Desc.)	5	3	60%	6	3	50%
Misc. other ideas	6	1	17%	4	3	75%
Total:	**158**	**140**	**89%**	-	-	-

5.1.3.2.5 Lesson 3 Exploration Activity 2 Question 3

This activity focused on an ice skater throwing a bowling ball (BB in Table 5.7). Students had been asked which object feels more acceleration when she threw the ball and in the present question were asked to explain their choice. The majority of students correctly recognized that the bowling ball had greater acceleration than the ice skater who threw it but their explanations for how they knew were varied. A plurality said they knew because it had less mass, which is true. Others noted that the bowling ball moves faster, and must therefore have accelerated more. This is also true. So this analysis presents the possibility of distinguishing between two sets of productive responses in addition to that are less productive. Mismatching in this case again was fairly uniformly distributed, and there were no clearly identifiable sources of mismatches from group to group.

This question also provides another example where it has been observed that reducing the number of groups could improve inter-rater agreement between computer and human. After this scheme had been developed, a three group scheme that focused entirely on the object was tested. That scheme produced inter-rater agreement above 92%, but the cost in terms of resolving the ideas students expressed in their reasoning was obviously too great to adopt that scheme. Mortality resulted in a relatively low number of responses to analyze for this question and whether a still better match rate could be achieved with more responses is a particularly tempting question for here (though it is clearly important throughout this analysis)

Table 5.7 Match/Mismatch details for lesson 3 exploration activity 2 question 3

Group	Number	Matched	Percent	Tagged	Mismatched	Percent
BB has less mass	32	29	91%	35	6	17%
BB has greater speed/veloc.	24	18	75%	23	5	22%
BB feels force (from skater)	18	11	61%	22	11	50%
Skater accelerates more	10	4	40%	7	3	43%
BB & Newton’s laws	9	5	56%	7	2	29%
Same acceleration	8	6	75%	6	0	0%
BB w/ other ideas	5	2	40%	5	3	60%
BB moves more/goes further	4	2	50%	5	3	60%
Total	**110**	**77**	**70%**	-	-	-

5.3 Discussion of Short Response Data and Analysis

The relatively large variation in the successful match rate between the trained computer model and the human coder suggests that this technique is not an unqualified success. It is unlikely that we can currently count on the training of computer models using existing routines such as SMO or NaïveBayes to classify student responses and generally expect agreement above 70-80%. At the same time the analysis has shown itself to be far better than random chance for every question we have applied the technique to. In five of the analyzed questions we see kappa statistics that are consistent with a reasonably good level of agreement. It is important to keep in mind that a human tutor will not correctly gauge student understanding and provide optimum feedback with 100% efficiency (Chi et al., 2004). The observations of Chi et al. were that tutors were only about 72% effective in gauging student understanding. This research was not performed in a physics context or with the same population so we cannot say that a real physics tutor would be wrong in gauging understanding 28% of the time, but this gives a far more realistic picture of what success looks like in the real human case. Therefore, while it is tempting to see a 25% failure rate with this automated approach and compare that to a human who always understands the students and always knows what to say to the student, that comparison is fallacious. In fact, no tutor always knows what a student understands or how to best reply and establishing a success criteria for our automated assessment approach is more a matter of establishing how well it needs to work to satisfy students need for feedback in practical implementation than a matter of striving for arbitrarily high standards of perfection. For questions that are lend themselves well to this approach we already have achieved results that are in the same range as Chi et al. observed for real human tutors, which gives us a very high incentive to continue work in this direction.

By looking at the group-by-group matching and mismatching it becomes clear that response sets that break down naturally into smaller numbers of groups the approach generally works better. There is some evidence that the method works best when the groups are similarly sized. Moreover we have demonstrated that even when the approach does do a good job of predicting our groupings across the board, we can typically see higher levels of agreement for groups that are more conceptually distinct. We observed 75% or better matching in at least two of the larger groups in every question we analyzed. At the same time we have not yet conclusively demonstrated that this approach will work well enough to provide students with feedback at a level that we could consider satisfactory for general use. While the method did correctly predict groupings effectively for five of the questions we analyzed, the issue of contamination, especially within larger groups is a potential cause concern. While improving the rate of agreement between human and computer would resolve this issue, simply because the total number of responses is fixed, we do not yet know if that is a reasonable expectation. Two possible approaches to increasing the rate of agreement are increasing the size of the data sets and performing additional modifications to the feature lists and trained models. The former has the potential to improve the results by virtue of simply training the models on more data and capturing more of the variable ways that students may express similar ideas. The later hinges on using human insight to focus the feature list to further exclude features that are not useful predictors, or to use harder rules (e.g. If a response contains term X, then *automatically* classify as Y, regardless of any other predictors) to result in better classification. Understanding what we can expect from scaling this approach up will help resolve possibility of using the first approach, and this warrants further research. If we significantly increase the number or responses analyzed, will we find the rate of agreement also increases or will we find more new ideas expressed that form still more small, and difficult to analyze groupings? More research is needed to answer that question. The other option, using human insight to focus the feature list, or using rules to better classify responses also warrants additional research, but already limitations are clear. With 150 responses generating feature lists of that are several hundred elements long, figuring out which ones are most important, relying on insight instead of the statistical analysis the extractor is already using is akin to looking for a needle in a haystack. A third approach, writing questions and activities with this analysis scheme in mind may be more productive. The activities and questions used in our system were not written with this scheme in mind. Instead it was discovered that the approach might have merit later as the project was progressing. We have observed that writing activities that naturally provide students with discrete physical and conceptual choices about which to construct their responses may provide response sets that are more amicable to this analysis approach. While this may appear in a sense to be a crouched form of multiple choice, the pedagogical benefits of asking students to answer questions freely, in their own words should not be overlooked. Observing that those responses do naturally break down into conceptually distinct groups additionally provides us with a way of processing large numbers of responses, as would be necessary when implementing a freely available online instruction system. That tract, coupled with collecting larger data sets to better understand how this approach scales, is likely the best avenue of development for furthering research and development into using the scheme for feedback in the type of learning environment we investigate.

Of greater concern for implementing this approach is the long development time and relatively high development effort that would be associated with each activity. In order to use this system for providing feedback it will be necessary to design the activity, collect a large number of student responses to activity questions (we are likely working at the bare minimum acceptable size of data set for this type of analysis), code those responses, train the models, and then implement the activity again with feedback. It is quite possible that an additional sequence of revision would be necessary. Doing this on a semester-by-semester basis (as one would likely have to do if collecting large numbers of responses from students actually enrolled in physics classes) would suggest a minimum development time of one year for a given activity. That is, the activity could be developed and a first set of responses collected in one semester, the data coded and models trained in time to implement the activity with feedback in the next semester. If it could be shown that this type of development cycle generally produced activities for which automated feedback was possible and that could be re-used for a number of years, this might justify the input effort. Clearly more research work is needed to understand whether it is possible to build activities with this analysis scheme in mind and have it reliably work.

5.4 Summary & Conclusions

In this chapter we have discussed the use of machine learning to train models to automatically classify student responses to short answer questions. Ultimately, this work shows initial promise, but also highlights the difficulties that must be overcome. High match rates for large groups look promising, but contamination brought on by poorer match rates in smaller groups can be problematic. Overall the match rates in the best cases are similar to the success rates in gauging student understanding that have been previously observed for real human tutors. The significant amount of effort required to analyze data in this way is concerning, however if the approach scales well that initial effort can be offset by time savings on subsequent assessment of student responses. We can conclude that the method has been demonstrated to be effective enough to warrant continued use in research contexts, and the prospects for use in instruction are quite encouraging, but additional work is required to understand how effectively the method can actually be implemented for providing students with feedback.

Chapter 6 - Conclusions

6.1 Overview of the work

The purpose of this project was to investigate how multimedia and natural language processing techniques could be used to scaffold a learning interaction that resembled tutoring in an online learning environment. This investigation is topical as online instruction is becoming increasingly common and the interactive multimedia capabilities that are routinely achievable on the Internet are increasingly sophisticated. We investigated these questions in the context of Newton's laws and focused our investigation on high school and college students studying conceptual and algebra-based physics. The system was constructed and the research was conducted between the summer of 2007 and the fall of 2011.

6.2 Research Questions Answered

In this project we sought to investigate the utility of an online learning environment that used natural language processing and video-based instruction to emulate tutoring. Our system focused both on conceptual understanding of physics as well as mathematical problem-solving. To pursue this line of investigation we conducted interviews with students who used our system, observed their use of the system on video and analyzed data collected from the system logs to obtain a picture of what is effective, what is not, and the reasons why. The central components of the system (SI tutor, video lessons and support multimedia) were investigated to the best that our data collection capabilities allowed and we have gained a significantly improved understanding of how this system should be developed if it is to become a viable means of instruction.

An important finding, that students did use the SI tutor as they worked with the lessons, is promising. At the same time evidence is conflicted about whether students view this interaction as asking a tutor questions. Evidence shows that students acknowledge the tutor as a person and discussing it as such. However, evidence also indicates that students treat it as they would a video player. Interview data suggest students have limited patience with poor performance on the part of the SI tutor and such experiences may spoil any suspension of disbelief that might be built up. The logs clearly indicate that students exhibit a preference for interacting with the tutor via menus, rather than by typing questions. This can be attributed to a combination of reluctance to spend time asking questions if they are unsure that the SI will have an answer to that specific question as well as uncertainty about what to ask at a given point in the tutoring session. The menu options solve both problems, but arguably make the interaction less like real tutoring. This result encourages us to ask the question of whether we should continue to focus on using the SI technology to promote a tutoring-based system. We discuss the advantages and disadvantages of doing so in more detail in section 6.3.

Analysis of the PALE logs suggest that students provided reasonably serious responses to most lesson questions, but struggled considerably with the video analysis, which was central to our lesson activities. Analysis of the PALE logs clearly indicates that a better video analysis interface is necessary for this system to be viable. Beyond the clear evidence of struggle in the logs, this aspect of the system was frequently brought up in interviews as a source of difficulty and frustration. There is significant evidence that students did not understand the video analysis either as analogous to an experiment or, more simply, as a means of using nature itself as a reference for checking our theoretical understanding. Instead the lack of precision generated anxiety in students who viewed these activities as something more like homework, which has right and wrong answers.

From a theoretical perspective, from empirical observations and based on pragmatic understanding of how tutoring works it became clear that formative feedback was an important feature of tutoring that our system did not provide. While providing automatic, computer-generated feedback to students on the correctness of a number is relatively easy, providing students with automatic feedback meaningful feedback on open-ended, conceptual questions is much more difficult. Unfortunately, it is also much more useful. An unfortunate truth of this kind of research is that it we *could not* provide that feedback within the system until we understood what students would say to our conceptual questions. Based on the responses students provided to our questions we investigated the application of cutting edge machine learning techniques to better allow the system to provide formative feedback to students, as a tutor would. We have seen both initial promise in that endeavor as well as clear indications of the challenges involved in developing that capability. Our machine learning analysis applied to manually coded data resulted in high levels of agreement for five out of the nine analyzed questions. We also observed much poorer results in the other four. Significant research efforts to better understand the range of utility for this approach are needed. At the same time these results represent one of the first applications of machine learning to the automated assessment of short-answer questions in physics education research. The method can be viewed as an enabling technology for much more sophisticated online homework systems for the very reason that we have chosen to use it.

Investigating the utility of the support multimedia suffered from a significant challenge in the form of establishing a clear signal that could indicate whether or not it was playing a useful role. Evidence from our interviews suggest that it was well-received by students, yet at the same time there is evidence that students had difficulty whether they had seen videos, images and what the contents of those images might have been. There is also evidence of students turning off support videos. Methods of better assessing the utility of this feature are discussed in section 6.3.

6.3 Future Work

In this section additional work that is suggested by this research or enabled by this research is discussed.

6.3.1 Modifications to PALE and further work with this system

Based on this work we can conclude that the following modifications to the PALE system can or should be made

- The video analysis interface must be improved and students must be informed on why connecting physics ideas to real experiments is important.
- Providing feedback to students on the correctness of their numerical results could help with their concerns about video analysis and improve their answers. Using the SI tutor to provide this feedback could also promote the SI as an important agent in the system.
- Further work on implementing a machine learning-based feedback system with the SI tutor should be done. Exploiting common right answers, wrong answers or physical focal points in the response sets shows enough promise to warrant further research. Again this may promote the SI tutor as an interactive agent in the system.

Beyond these modifications it would be desirable to collect more student responses to the questions in our lesson activities, particularly ones that featured questions that could be well-grouped. This work would give us valuable information about how the machine learning approach to analyzing the response sets scales. Understanding whether we can expect significantly better matching if we work from 300 responses, or whether we would need significantly more than that to obtain acceptable match rates for providing students with feedback.

6.3.2 Further Work

Beyond the investigations that relate to direct modifications to the PALE that are proposed in the previous section there are less immediate research directions that that can be pursued from our current position.

6.3.2.1 Fully-automated clustering of student responses

To begin with there are other approaches to data mining that could be used to analyze our student response data. The approach we used here has a serious drawback in that someone has to code a large corpus of data before the method can be applied. Other data mining approaches can be used to automatically cluster text elements in a data corpus. A good introduction is provided by Tan, Steinbach and Kumar in Chapter 8 of *Introduction to Data Mining* (2006). Many of these approaches do not require training with user-coded data. This would be a considerable improvement. Initial tests using several common clustering algorithms did not provide conceptually meaningful clusters, but there is a wide range of approaches and our testing was not exhaustive. The goal at that time was to establish an approach that could successfully be applied in our situation. The procedure described produced better results faster and was chosen for this research. A natural question to ask is whether we can find an automatic clustering algorithm that does a good job of reproducing the clusters we have already discovered in this data. That result might give us better insight into which, if any, of the many clustering algorithms might be appropriate for analyzing this kind of data and how to best apply them to obtain meaningful results without manually grouping several hundred responses.

6.3.2.2 Identifying student pathways through lesson materials

Discovering useful groupings that can be applied to student responses to lesson questions provides us with potential for exploring the efficacy of the tutoring system in a more dynamic way than a pre-test/post-test design allows. Once we have clearly established groupings of student responses we can ask an interesting question- What is the probability that a student's response to question 2 is in a certain group, given that his or her responses to question 1 was in a known group? By examining successive questions for a reasonably large number of students we can begin to establish what we might call "conceptual pathways" through the series of lessons. We have reason to believe naturally emergent groupings of student responses will tell us, first something about how students were thinking when they answered that question, but also about how they think about relevant physics concepts in general. Since there are connections between the physics concepts addressed by different questions, even in different activities, we would expect correlations across multiple questions and multiple activities to emerge. In order to investigate these pathways this stage of analysis should focus on data from students who have completed the entirety of the lesson materials. Contingency tables are a standard statistical approach to looking for relationships between two categorical variables (Agresti, 2009). Analyzing more than two categorical variables, as we have in this case, is commonly done using Log-linear analysis, which is a form of multi-variable regression modeling (Fienberg, 2007). This analysis would be particularly interesting if the groupings of responses that we identify can be shown to have conceptual connections across questions that enable us to meaningfully analyze the responses to one set of responses using groupings discovered via another. While it would be pure speculation to suggest that this result is likely or unlikely, it is an interesting possibility that could emerge and that should be investigated.

6.3.3 Understanding this Work in the Context of Emerging Internet Technology

One of the interesting and challenging aspects of this project is the rapidity with which the technology has advanced while the project was completed. It is hard to imagine, but true that prior to February 2005 there was no such thing as YouTube. Prior to February of 2004 there was no Facebook. These websites and many others that are similar, and that characterize the new and important ideas of user-developed content and experiences, really began to emerge as some of the most visible and important players in our culture at the same time that this project began its development. These types of websites have evolved in symbiotic relationship with hardware technologies that feed into their capabilities. These technologies include tablets, cellular telephones, and inexpensive digital cameras (static and video), which enable users to generate content anywhere at any time, publish that content to the web, respond to other's content, and interact in asynchronously or in real-time. It is therefore not an overstatement to say that the world that our system was designed to fit into is very different from the world that it now inhabits. The critical change that is occurring is that we are transitioning towards an Internet that provides people with more power in creating their own content (text, music, images, and video) and which is extremely powerful and socially interactive in the dissemination of that content. The PALE system was designed based on the idea that the Internet could be used to provide interactive instruction via a computerized interface, and we have constructed a proof-of-concept that demonstrates that in many ways this is possible. Furthermore we have identified clear pathways towards improving upon our system. However, it is important to pause at this moment and reflect on the technological landscape that our improved system would occupy and examine where a system like this fits in.

Based on the observation that students use menus to interact with the SI tutor far more frequently than typing questions it is valid to ask about alternatives to further development of the system as a tutor. It is possible to cast the SI video interface either as a tutor or as something more like a video FAQ. The development effort required to generate and refine the question list and natural language processing software needed to answer typed questions is considerable. If students mostly use the menus, rather than typing questions, then this effort may not make sense. It is possible that introducing the modifications discussed previously that will increase the SI's interactivity will also encourage students to ask questions rather than selecting them, and this should be investigated, but at the same time it is worth investigating the utility of a system like this that is only menu-driven. A system like that would not aspire to emulate tutoring, but instead should be considered video-supported lesson activities. A logical course of action would be to develop an online system that is more flexible. One that would allow teachers to load their own lesson questions, lesson video clips, and any explanatory support that they would like to generate activities and collect their students' responses. A system like this could distribute the task of generating and testing activities over many teachers rather than a few researchers. At the same time, if teachers were willing to share their activities one could envision having banks of hundreds of student responses for analysis by researchers. In essence this is a two roads forward picture. On the one hand further development and testing should be done to determine how interactive the SI tutor can be made. On the other hand a system that does not rely on the complex natural language processing can open up the development of activities to a wider population of instructors and researchers. When this project was first proposed, it would not have made any sense to talk about crowd-sourcing the development of video-based online learning modules. Since this project was proposed YouTube, Facebook and a variety of other utilities that thrive on user-generated content have changed the landscape. Inexpensive video cameras the connect directly to computers and a wealth of public domain and creative commons licensed video means that the average high school or college teacher can make activities like the ones we have made and publish them on the Internet. Some teachers are already using homemade video or video analysis in their classrooms. YouTube has become a common resource for physics teachers (Riendeau, 2012). Establishing a centralized forum for that which promotes sharing could easily result in more quickly and accurately identifying of good activities and the kinds of explanatory support that students might typically need for each activity. It is possible, but not necessary to feed the results of this line of development back into the other and let researchers use what teachers and students have discovered about the activities to build interactive video tutors. Taking the development in this direction could mitigate the single most challenging aspect of this research: the long development cycle.

6.4 Final Conclusions

While there is great interest in human-computer tutoring prior to the development of our system, the combination of natural language processing and video-based instruction has not been explored. Our system is the first, to the best of our knowledge, that uses video of actual human tutors to provide that tutoring interface and combines that feature with natural language processing instead of artificial intelligence. The combination of these two features provides a clear pathway towards a tutoring system that is conversational in nature and that resembles video conferencing. This SI-tutor has shown initial promise as an interactive means of instruction, with students querying the tutor to obtain information at a rate that is much higher than observed in classroom environments (Graesser & Person, 1994).

There is evidence that students may use the tutor without recognizing conflict between their current understanding at the information presented by the tutor. This is concerning , but at the same time, our analysis has revealed the importance of providing students with relevant feedback that is specific to their actions while working with the system. This is in agreement with several theoretical models of tutoring. It is therefore possible that the students can be provided with feedback in a way that encourages them to face this conflict. It is tempting to ask why the system was designed without this feedback mechanism, but it is important to note that providing automated feedback to students without the data that provides information about how they will respond to lesson questions is extremely difficult, if not impossible. It was necessary to obtain student responses to lesson questions before feedback could be introduced. In a sense this is an important result of this project, which unfortunately cannot easily generalize. In order to explore analysis schemes that cab provide feedback in the future we must first collect data with a system that cannot provide feedback. Unless further research reveals general principles about how students respond to these types of activities, it is likely that future efforts will have to develop in the same way. It is much harder to predict how students will respond to questions than it is to observe them answering them. Even without the ability to generalize, this approach shows significant promise for providing feedback on an activity-by-activity basis. Our attempt to exploit this type of analysis scheme has provided us with insight into how to better create lesson activities with the kinds of distinct physical elements and salient ways of conceptualizing physical behavior that promote success with this approach. While the approach has not worked perfectly some success has clearly been demonstrated. Furthermore it is important to remember that real human tutors will not accurately gauge student understanding all the time and so the real value of this demonstration may be higher than one might initially infer from a 70-80% inter-rater agreement. This result points us clearly in a direction of increasing the interactivity of our system and allowing it to function more like an actual tutor.

One of the most important questions that remains unanswered is how this approach will scale. It is expected that the automated assessment scheme will work better with larger data sets, but this must be demonstrated in the future. One of the challenges associated with this project was establishing a clear metric for distinguishing the relative educational benefits of the different multimedia that supported the SI tutor. While we do not see clear evidence for the superiority or inferiority of the different multimedia support that we used with our system we have identified additional research methods that may provide us with more insight into the relative merits of these modes of presentation in the future.

While we observed that video measurement in this context provided a significant stumbling block for students working with the lesson materials, we have identified methods of improving that facet of the system. Students indicated that they liked the videos as they did connect physics concepts with real life and provided a live-action demonstration of the concepts. An interesting finding, that students framed the lessons as a form of homework activities is important to note because the inherent uncertainty in measurement becomes a potential source of frustration or concern for students who are used to numbers being provided in their homework, and being provided with essentially infinite precision. While this may be viewed as a burden, it may be possible to turn it around and look for ways to use this situation to encourage students to think about uncertainty in science on a more regular basis without getting into the complexities of error analysis. This becomes an intrinsic way of getting students used to the idea that all numbers are somewhat uncertain.

Looking beyond the immediate focus of our project, we have pointed out that this type of system was devised at a point in time when the Internet was quite different, and it is important to consider the direction that emerging technology is headed when deciding on a course for future research. Even if continuing work on this type of tutoring system is not deemed to be of continuing interest, the results of our research may be helpful in informing research in online learning environments that look quite different. Our work in automated assessment of short answer text responses points towards an important area that is of interest to anyone developing online learning systems, or online homework services. This is an area in which we are one of the earliest explorers in physics education research. Additionally we have discussed the potential of exploiting the explosion in user-generated (and published) multimedia to build a flexible framework of a system similar to ours that would allow students and teachers to build their own interactive multimedia learning experiences. While we have clearly laid out a way forward in developing an interactive tutoring system, we have also made significant contributions to the exploration of how to build, and deploy interactive multimedia-based instructional experiences. We have exploited the Internet's ability to access geographically separated audiences to collect and analyze relatively large amounts of data and then mined that data for themes and analyzed in other ways to gain insight into the relative merits of different facets of the experiences. This type of research can be viewed as a cycle of development and ultimately there are considerable opportunities for additional research exploring multimedia in online instruction and the use of natural language processing and machine learning for improving the interactivity of online learning environments.

References

Agresti, A. (2009). *Statistical Methods for the Social Sciences*, Upper Saddle River, New Jersey: Pearson Prentice Hall.

Atkin, J. M., & Karplus, R. (1962). Discovery or Invention, *Science Teacher*, 29(5), 45

Bloom, B. S. (1984). The 2-sigma problem: The search for methods of group instruction as effective as one-to-one tutoring, *Educational Researcher*, 13(6), 4-16.

Beichner, R. (1996). The impact of video motion analysis on kinematic graph interpretation skills, *American Journal of Physics*, 64 (10), 1272-1277.

Beilin, H. (1992). Piaget's Enduring Contribution to Developmental Psychology, *Developmental Psychology*, 28(2) 191-204.

Brown, D. & Cox, A. J. (2009). Innovative uses of video analysis, *The Physics Teacher*, 47(3), 145-150.

Brungardt, J. & Zollman, D. A. (1995). The influence of interactive videodisc instruction using simultaneous-time analysis on kinematics graphing skills of high school physics students, *Journal of Research in Science Teaching*, 32(8), 855-869.

Butcher, P. G., & Jordan, S. (2010). A comparison of human and computer markings of short free-text student responses. *Computers and Education*, 55(2), 489-499.

Cadmus, R. R. (1990). A video technique to facilitate the visualization of physical phenomena, *American Journal of Physics*, 58(4), 397-399.

Champagne, A., Klopfer, L., & Anderson, J. (1980). Factors influencing the learning of classical mechanics, *American Journal of Physics*, 48(12), 1074-1079.

Chi, M. T. H., Siler, S. A., & Jeong, H. (2004). Can tutors monitor students' understanding accurately?, *Cognition and Instruction*, 22(3), 363-387.

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

Chi, M. T. H., De Leeuw, N., Chiu, M. H., & LaVancher, C. (1994). Eliciting self-explanation improves understanding, *Cognitive Science*, 18, 439-477.

Christel, M., Kanade, T., Mauldin, M., Reddy, R., Sirbu, M., Stevens, S. & Wactlar, H. (1995). Informedia Digital Video Library, *Communications of the ACM*, 35(4), 57-58.

Clark, R. E. (1983), Reconsidering research on learning from media, *Review of Educational Research,* 53(4), 445–459.

Clark, R. E. (1994), Media will never influence learning, *Educational Technology, Research and Development*, 42(2), 21–29.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37-46.

Cohen, J. (1977). *Statistical Power and Analysis for the Behavioral Sciences*, (Rev. ed.). New York: Academic Press.

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

Cresswell, J. W. (2007). *Qualitative Inquiry and Research Design Choosing Among Five Approaches,* London, UK: Sage Publications.

D'Mello, S. K., Dowell, N. & Graesser, A. (2011). Does it really matter whether students' contributions are spoken versus typed in an intelligent tutoring system with natural language? *Journal of Experimental Psychology-Applied*, 17(1), 1-17.

Dufresne, R. J., Gerace, W. J., & Leonard, W. J. (1997). Solving Physics Problems with Multiple Representations, *The Physics Teacher*, 35(5), 270-283.

Ellson, D. G. (1976). *The Psychology of Teaching Methods,* Chicago: University of Chicago Press.

Escalada, L. T., & Zollman, D. A. (1997). An investigation on the effects of using interactive digital video in a physics classroom on student learning and attitudes, *Journal of Research in Science Teaching*, 34(5), 467-489.

Festinger, L. (1957). A Theory of Cognitive Dissonance, Stanford CA: Stanford University Press.

Fienberg, S. (2007). The Analysis of Cross-Classified Categorical Data, New York, New York: MIT Press

Fitz-Gibbon, C. T. (1977). *An analysis of the literature of cross-age tutoring,* Washington D.C: National Institute of Education.

Flavell, J. H. (1996). Piaget's Legacy. *Psychological Science*, 7(4), 200-203.

Flavell, J. H. (1992). Cognitive development: Past, present and future. *Developmental Psychology, 28, 998-1005.*

Freeman, M. A., Hennessy, E. V., and Marzullo, D. M. (2001). Defensive evaluation of antismoking messages among college-age smokers: The role of possible selves, *Health Psychology*, 20(6), 424-433.

Fuller, R. G., Campbell, T. C., Dykstra, D. I. & Stevens, S. M. (2009). *College Teaching Development of Reasoning*. Charlotte, North Carolina: Information Age Publishing.

Graesser, A. C., Jeon, M., & Dufty, D. (2008). Agent technologies designed to facilitate interactive knowledge construction, *Discourse Processes*, 45, 298–322.

Graesser, A. C. & Person, N.K. (1994). Question asking during tutoring, *American Educational Research Journal*, 31(1), 104-137.

Hsu, L. & Heller, K. (2009). Computer Problem-Solving Coaches, *Proceedings of the NARST 2009 annual meeting,* April 17-21, Garden Grove, CA.

Hammer, D., Elby, A., Scherr, R. E., & Redish, E.F. (2005). Resources, framing and transfer in J. Mestre (Ed.) *Transfer of Learning: Research and Perspectives*, (pp. 89-119) Information Age Publishing: Greenwich, CT

Hammer, D. (2000). Students resources for learning introductory physics, *American Journal of Physics, Physics Education Research Supplement*, 68(SI), S52-S59.

Hestenes, D., Wells, M. & Swackhammer, G. (1992). Force Concept Inventory, *The Physics Teacher*, 30, 141-158.

Holton, G., Rutherford, F. J., & Watson, F. G. (1971). *About The Project Physics Course*. New York : Holt, Rinehart & Winston Inc.

Hundeide, K. (1977). *Piaget I kritisk lys.* Trondheim: Cappelens (Piaget in a Critical Light)

Inhelder, B. & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence.* New York: Basic Books.

Jordan, S. (2012). Student engagement with assessment and feedback: Some lessons from short-answer free-text e-assessment questions *Computers & Education*, 58, 818-834.

Karplus, R, & Butts, D. P. (1977). Science teaching and the development of reasoning, *Journal of Research in Science Teaching*, 14(2), 169-175.

Kanim, S. E., & Subero, K. (2010). Introductory labs on the vector nature of force and acceleration, *American Journal of Physics*, 75(5), 461-466.

Kearney, M. (2004). Classroom use of multimedia-supported predict-observe-explain tasks in a social constructivist learning environment, *Research in Science Education*, 34, 427-453.

Kim, S. B., Han, K. S., Rim, H. C., & Myaeng, S. H. (2006). Some Effective Techniques for Naïve Bayes Text Classification, *IEEE Transactions on Knowledge and Data Engineering,* 18(11) 1457-1466.

Krippendorff, K. (1980). *Content analysis: an introduction to its methodology*, 1st ed. Sage Publications, Thousand Oaks, London

Landis JR, Koch GG (1977). The measurement of observer agreement for categorical data. *Biometrics* 33:159–174

Laws, P. & Pfister, H. (1998). Using digital video analysis in introductory mechanics projects, *The Physics Teacher*, 36, 282-287.

Larkin, J. H. (1983). The role of problem representation in physics, in D. Gentner & A. L. Stevens (Eds.) *Mental Models*, (pp.75-98), Hillsdale, New Jersey: Lawrence Erlbaum Associates Inc.

Lewis, R. A. (1995). Video introductions to laboratory: students positive, grades unchanged, *American Journal of Physics*, 63(5), 468-470.

Madsen, A. M., Larson, A. M., Loschky, L. C., & Rebello, N. S Differences in visual attention between those who correctly and incorrectly answer physics problems, *Physical Review Special Topics: Physics Education Research (To be published)*

Marinelli, D., & Stevens S. (1998). Synthetic Interviews: The art of creating a "dyad" between human and machine-based characters. *4th IEEE Workshop on Interactive Voice and Technology for Telecommunications Applications*, Torino, Italy.

Marton, F. (1981). Phenomenography- Describing conceptions of the world around us. *Instructional Science*, 10, 177-200.

Marton, F. (1988). "Describing and Improving Learning" in R.R. Schmeck (Ed.) *Learning Strategies and Learning Styles*, (pp.53-82). New York: Plenum Press.

McDermott, L.C., Rosenquist, M. L., & van Zee, E. H. Student difficulties in connecting graphs and physics: Examples from kinematics, *American Journal of Physics,* 55(6), 503-513.

McKinnon, J. W., & Renner, J. W. (1971). Are colleges concerned with intellectual development?, *American Journal of Physics* 39(9), 1047-1052.

Mayer, R.E. (2005). Cognitive Theory of Multimedia Learning. In R. E. Mayer (Ed.), *The Handbook of Multimedia Learning* (pp. 31-48). New York: Cambridge University Press.

Mestre, J.P., & Tougher, J. (1989). Cognitive research- what's in it for physics teachers? *The Physics Teacher*, 27, 447-451.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information, *Psychological Review*, 63(2), 343-355.

Nakamura, C. M., Murphy, S. K., Zollman, D., Christel, M., & Stevens, S. (2011). Finding Meaningful Search Features for Automated Analysis of Short Responses to Conceptual Questions in Rebello, N. S., Englehardt, P. and Singh, C. (Eds.) *The Physics Education Research Conference Proceedings-2011*, (pp. 283-286) AIP Conference Proceedings 818 American Institute of Physics, Melville, NY 2012

Nakamura, C. M., Murphy, S. K., Zollman, D., Christel, M., & Stevens, S. (2010). Pilot Testing of the Pathway Active Learning Environment in Singh, C., Sabella, M. and Rebello, N. S. (Eds.) *The Physics Education Research Conference Proceedings-2010,* (pp. 237-240) AIP Conference Proceedings 818 American Institute of Physics, Melville, NY, 2011

Nakamura, C. M., Murphy, S. K., Juma, N. M., Rebello, N. S., & Zollman, D. (2009). “Online Data Collection and Analysis in Introductory Physics” in Henderson, C., Sabela, M. and Singh, C. (Eds.) *The Physics Education Research Conference Proceedings-2009*, (pp.217-220) AIP Conference Proceedings 818 American Institute of Physics, Melville, NY, 2010

Nehm, R. H., Haertiz, H. (2011). Human vs. computer diagnosis of students’ natural selection knowledge: Testing the efficacy of text analytic software, *Journal of Science Education & Technology*, DOI: 10.1007/s10956-011-9282-7

Neuschatz, M. & McFarling, M. (2000). Background and professional qualifications of high-school physics teachers, *The Physics Teacher*, 28(2), 98-104.

Noble, L., Zollman, D., & Satern, M. (1988). Physics of Sports: An Interactive Videodisc for Analyzing the Motion of Athletes, *Proceedings of the 6th ISBS: Sports Biomechanics.*

Okita, S. Y., Bailenson, J. & Schwartz, D. L. (2007) “The mere belief of social interaction improves learning,” *Cognitive Science Conference*

Paul, J. P., diSessa, A. A., & Roschelle, J. (1994). Misconceptions reconceived: A constructivist analysis of knowledge in transition, *Journal of the Learning Sciences*, 3(2), 115-163.

Page, A., Moreno, R., Candelas, P., & Belmar, F. (2008). The accuracy of webcams in 2D motion analysis: sources of error and their control. *European Journal of Physics*, 29, 857-870.

Pearlstein, S. (2011, May 28). Mark them Tardy to the Revolution. *The Washington Post*, Retrieved from http://www.washingtonpost.com/steven-pearlstein-mark-them-tardy-to-the-revolution/2011/05/24/AG1vKYDH_story.html

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

Phillips, D. C. (1995). The good, the bad and the ugly: The many faces of constructivism. *Educational Researcher*, 24(7), 5-12.

Platt, J. C. (1998). Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines, *Microsoft Research Technical Report (Unpublished).*

Redish, E. F. (1994). Implications of cognitive studies for teaching physics, *American Journal of Physics,* 65(1), 796-803.

Redish, E. F. (2003). *Teaching Physics with the Physics Suite*: Wiley.

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

Richtmyer, F. K. (1933). Physics is physics, *The American Physics Teacher*, 1(1) 1.

Reeves, B. & Nass, C. (1996). *The Media Equation*, Stanford: CA, Center for the Study of Language and Information.

Riendeau, D. (2012). YouTube Physics, *The Physics Teacher*, 50(4), 249.

Rosé, C., Wang, Y.C., Cui, Y., Arguello, J., Stegmann, K., Weinberger, A., & Fischer, F. Analyzing collaborative learning processes automatically: Exploiting the advances of computational linguistics in computer-supported collaborative learning. *International Journal of Computer-Supported Collaborative Learning*, (2007). DOI: 10.1007/s11412-007-9034-0

Rosenshine, B., & Furst, N. (1969). *The effects of tutoring upon pupil achievement: A research review*, Washington D.C.: Office of Education.

Scherr, R. E., & Hammer, D. (2009). Student behavior and epistemological framing: Examples from collaborative active-learning activities in physics, *Cognition and Instruction*, 27(2), 147-174.

Schwartz, D. L. & Bransford, J. D. (1998). A time for telling, *Cognition and Instruction*, 16(4), 475-522.

Smith, A.D., Mestre, J. P., & Ross, B. H., (2010). Eye-gaze patterns as students study worked-out examples in mechanics. *Physical Review Special Topics: Physics Education Research*, 6(2), 020118.

Stelzer, T., Gladding, G., Mestre, P., & Brookes, D. (2009). Comparing the efficacy of multimedia modules with traditional textbooks for learning introductory physics topics, *American Journal of Physics,* 77, 184-190.

Stevens S., Zollman, D., Christel M., & Adrian B. (2007). Virtual Pedagogical Agents as Aids for High School Physics Teachers, *International Conference on Interactive Computer-aided Learning.*

Sweller, J. (2005). Implications of Cognitive Load Theory for Multimedia Instruction. In R.E. Mayer (Ed.), *The Handbook of Multimedia Learning* (pp.19-30). New York: Cambridge University Press.

Tan, P. N., Steinbach, M. & Kumar, V. (2005). Introduction to Data Mining, Boston: Addison Wesley

Trowbridge, D. E., & McDermott, L. C. (1981). Investigating student understanding of the concept of acceleration in one dimension, *American Journal of Physics*, 49(3), 242-253.

Trowbridge, D. E., & McDermott, L. C. (1980). Investigating student understanding of the concept of velocity in one dimension, *American Journal of Physics*, 48(12), 1020-1028.

VanLehn, K., Lynch, C., Schulze, K., Shapiro, J. A., Shelby, R., Taylor, L., Treacy, D. Weinstein, A., & Wintersgill, M. (2005). The Andes Physics Tutoring System: Lessons Learned, *International Journal of Artificial Intelligence in Education*, 15(3), 147-204.

VanLehn, K., Siler, S., Murray, C., Ymauchi, T., & Bagget, W. B. (2003). Why do only some events cause learning during human tutoring? *Cognition and Instruction*, 21(3), 209-249.

Vygotsky, L. S. (1997). *Educational Psychology*. Boca Raton: St. Lucie Press.

Vygotsky, L. S. (1978). *Mind in Society: The Development of Higher Psychological Processes*. Boston: Harvard University Press.

White, R. T., & Gunstone, R. F. (1992). *Probing Understanding*, Great Britain: Falmer Press.

Wood, D., Bruner, J. S., & Ross, G. (1976). The Role of Tutoring in Problem Solving. *Journal of Child Psychology and Psychiatry*, 17, 89-100.

Zollman, D., & Fuller, R. G. (1994). Teaching and Learning Physics with Interactive Video, *Physics Today*, 47(4), 41-47.

Zollman, D. (1990). Learning cycles for a large enrollment class, *The Physics Teacher*, 28, 20-25.

Appendix A - PALE Lesson Materials

Lesson 1: Newton's First Law

<u>Section 1: Exploration</u>

Ex1: Motion on a Low-Friction Track

(Key Idea: The ball maintains its speed as it moves along the track)

Directions: Use the video of the ball rolling along the track to answer the questions.

1. What is the speed of the ball near the beginning and end of the track?
2. How did you calculate the speed of the ball?
3. Does the ball's speed change significantly from the beginning to the end?
4. What would you predict the ball's speed to be somewhere in the middle of the track? Why?
5. What would you predict the speed of the ball to be at the end of a track that is twice as long? Why?

Ex2: Coffee Cup Video

(Key Idea: The cup resists change in motion and simply falls when the car moves)

Directions: Use the video of the car and coffee cup to answer the questions.

1. What is the initial speed of the coffee cup and car?
2. When the car starts to move does the coffee cup's velocity in the horizontal direction also change?
3. How can you tell?
4. Why do you think the cup behaves as it does?

Ex3: Car Crash Test

(Key Idea: The dummy maintains its speed until the dash board exerts a force on it to bring it to rest)

Directions: Use the car crash test video to answer the questions.

1. Record an estimate of the speed of the car and the dummy before and after the crash.
    a. Car before:
    b. Dummy before:
    c. Car after:
    d. Dummy after:

2. The car's speed is roughly constant prior to the crash. How does the speed of the dummy after the car stops moving compare to the car's original speed?

3. What, if any, forces act on the dummy?

4. Do any of the forces you identified in question 4 affect the dummy's horizontal motion?

5. Why does the dummy finally stop moving?

Section 2: Concept Introduction

Let's look at what can be learned from the exploration activities you've just completed.

In the first activity you watched a ball roll across a flat track. You should have found that the ball's speed doesn't change much from the beginning to the end of the track. If the ball's speed is the same at the beginning and end you might think it should be the same in the middle. Watching the video gives no indication that the ball speeds up or slows down, so you'd be right. This tells us something important: If an object is moving with some velocity, then it will continue moving with that same velocity unless something pushes or pulls on it to change its velocity. For example if we put a block in front of the ball we know it'd bounce off, and its velocity would change. So it's safe to guess that if we double the length of the track then the ball's speed at the end would still be the same as before. At the same time, we know from our life experiences that if the track is long enough this will stop being the case. If the track were ten times longer, then the ball might begin to slow down visibly due to friction. The track pulls on the ball, causing its velocity to change.

In the second activity you looked at a coffee cup on the back of a car as the car drove off. Here you saw that even when the car began to move the cup wanted to stay in one place. Its velocity in the horizontal direction didn't change. You can see this by looking at the cup's position with respect to the red line. Ultimately it falls down because gravity exerts a force pulling the cup downward. This tells us something else that's important: constant velocity can mean zero velocity. Therefore if an object is sitting still, it will continue to sit still unless something pushes or pulls on it to make it move. In this case the surface of the car and the bottom of the cup are smooth and there is little friction between them, so the cup doesn't feel much horizontal force.

In the third activity you looked at a crash test in which a crash test dummy is not wearing a seatbelt. You estimated the velocity of the car and dummy before and after the crash. The first thing to note is that the car and dummy move together before the crash so whatever speed one has, the other has the same speed. Next note that after the crash the car is not moving, so its speed after the crash is zero. Then, all that remains is the speed of the dummy after the crash. A careful estimation (or even just rough observation) shows that the dummy has roughly the same speed after the crash as it had beforehand, until it collides with the dashboard. The dashboard exerts a force on the dummy, stopping it. Again this shows us Newton's first law. The dummy is moving with constant velocity and continues to do so until something (the dashboard) exerts a force on it to bring it to rest.

In each of these three videos we see the same behavior: objects will continue to move with constant velocity unless an external force acts to change that velocity. This is Newton's First law of motion. If you don't understand anything I've said here, you might want to ask me about it.

<u>Section 3: Application</u>

<u>App1: Coin and Graduated Cylinder</u>

(Key Idea: You must bring the coin into motion, and then stop the cylinder. The coin will maintain its motion and thus fly out of the cylinder)

Directions: Answer the first two questions BEFORE viewing the video clip. Answer the second two questions AFTER viewing the video clip.

1. Consider a coin stuck lightly to the bottom of a graduated cylinder. The cylinder is narrow and long so you can't reach in and get it. Given your knowledge of Newton's first law, how would you go about getting the coin out of the cylinder?

2. Explain your answer completely and clearly.

3. Were your answers to the previous two questions correct?

4. Explain how the depicted solution works in terms of Newton's 1st law.

App2: Liquid Filled Carts

(Key Idea: The fluid will resist changes in motion and therefore collect at the back of the container when it begins to move.)

Directions: Answer the first question BEFORE viewing the video. Answer the second question AFTER viewing the video.

1. Consider a cart with wheels that has a fluid-filled container on it. Think about what would happen if we pulled on it with a constant force. Which answer best describes what the surface profile of the fluid will look like once we pull on the cart?

a) The fluid surface will be flat.
b) The fluid surface will slope such that it is lower near the front (where the string connects).
c) The fluid surface will slope such that it is higher near the front (where the string connects).
d) The fluid surface will slope such that it is low in the middle and high on the ends.
e) None of these.

2. Explain your answer to question 1.

3. Now, watch the video. Was your answer to question 1 correct?

4. Explain why the profile might look the way it does using Newton's first law.

App3: Coin and Beaker

(Key Idea: The coin resists changes in motion and then falls directly into the beaker when the card is removed)

Directions: Answer the first two questions BEFORE viewing the video. Answer the last two questions AFTER viewing the video.

1. In the video a coin rests on a card that is sitting on a glass beaker. Once the video is started the card will be pulled quickly. Which of the choices below best describes the trajectory of the coin after the card is pulled?

   a) The coin will fall down and towards the right landing in the beaker.
   b) The coin will fall down and towards the left landing in the beaker.
   c) The coin will fall straight down into the beaker.
   d) The coin will move with the card and will not fall at all.
   e) The coin moves but doesn't land in the beaker.

2. Explain your answer to question 1.

3. Now, watch the video. Was your answer to question 1 correct?

4. Explain the trajectory of the coin using Newton's first law.

Lesson2: Newton's Second Law

Section 1: Exploration

Ex1: Pulling a Mass with Constant Force

(Key Idea: Increasing the force applied to the cart increases the acceleration. Increasing the mass of the cart decreases the acceleration.)

Directions: Use the video clips to answer the questions below.

1. Measure the acceleration of the cart in each of the three videos. You should use the ruler in the video and the fact that the time between consecutive frames is 0.03s. Remember that for an object that starts from rest and undergoes constant acceleration the displacement is given by $\Delta x = 1/2 \cdot a \cdot t^2$. It is easiest to use this relationship to get the acceleration.
    a. Video 1: Acceleration (m/s^2)
    b. Video 2: Acceleration (m/s^2)
    c. Video 3: Acceleration (m/s^2)

2. How did you use the information provided in question 1 to determine the acceleration in each video? Please explain the process you used to find the acceleration and not repeat the instructions in question 1.

3. What is the difference in the acceleration when the applied force was roughly doubled (compare video 1 and video 2)?

4. What is the difference in the cart's acceleration when its mass was doubled (compare video 2 and video 3)?

5. What is the simplest relationship you can infer between force, mass and acceleration from your observations?

Ex2: Impulsive Forces versus Constant Forces

(Key Idea: When you stop applying a non-zero net force to an object the object stops accelerating.)

Directions: Use the two videos to answer the questions below.

1. Does the puck accelerate at any point during the first video clip?

2. If so, when does the acceleration stop and how can you tell?

3. Does the puck accelerate at any point during the second video clip?

4. If so, when does the acceleration stop and how can you tell?

Ex3: Net Force is the Sum of all Applied Forces

(Key Idea: It is the net force that causes objects to accelerate.)

Directions: Use the video clip to answer the questions below.

In the video we see three configurations in which two masses are connected to a string hung over a pulley. The pulley has very little friction on its axel. In the first two configurations the masses are equal (0.5kg each). In the third configuration additional mass has been added to the right side (the total mass is 0.550kg). Answer the following questions based on what you see.

1. When the two masses hang at the same height, what can exert forces on each mass? Do the masses feel the same forces?

2. What is the total force on each mass?

3. When the two masses hang at different heights, but are still motionless, what can exert forces on each mass? Do the masses feel the same forces?

4. What is the total force on each mass?

5. In the third demonstration look at the mass on the right. How does its acceleration compare to the acceleration an equal mass would feel if it were just dropped? Why do you think that might be?

Section 2: Concept Introduction

(Spoken by SI or read by student)

Let's look at what can be learned from the exploration activities you've just completed.

In the first activity you were asked to find the acceleration of three carts and the easiest way to do that was to use the kinematic relationship d=1/2at^2. You could solve for the acceleration, and use the video clip to find the displacement and time interval. We can use the first video to establish a reference for comparison for the other two. We applied a force to a cart and measured its acceleration. Remember, a force is just a push or a pull, so we pulled on the cart. We then changed the force that we applied to the cart, and then changed the mass of the cart. What you hopefully saw was that doubling the mass of the cart roughly halved the acceleration it felt. At the same time doubling the force applied to the cart roughly doubled its acceleration. Scientifically this is not enough observation to establish a relationship between force, mass and acceleration, but others have done much more observation and we can compare our results with theirs. Their observations are that the net force is related to the mass of an object and its acceleration, by the equation F=ma. In this experiment we indeed saw that keeping the force fixed, and doubling the mass resulted in the acceleration being halved. Similarly keeping the mass fixed and doubling the force results in the acceleration being doubled as well. So although we haven't proven F=ma our observations definitely support it. This is what Newton's second law says: the net force felt by an object is equal to its mass multiplied by its acceleration. Remember that acceleration is the rate at which an object's velocity is changing in time, so Newton's second law tells us how forces change an object's velocity.

In the second activity you were asked to look at two pucks exhibiting two different kinds of motion. In the first video the puck was accelerated by a constant force that was provided by tilting a frictionless air-table. In the second video the puck was accelerated by an instantaneous force that was provided by a launching device. In this activity the key question is whether each puck feels a force. The answer is yes, but with important differences. The first puck feels a constant force that acts over the course of its entire motion. As a result we see that over constant time intervals its displacement is constantly increasing. This indicates that its velocity is changing; it is accelerating. The second puck, however feels a force only while it is in contact with the launcher. During that time it accelerates because its initial velocity is zero, and its final velocity is not, but once it leaves the launcher it moves with the same displacement in every time interval, so its velocity is constant. This shows us that an object's velocity only changes, that is the object accelerates, while the force is actually acting on it.

Unlike in the first two videos, the masses in the third video obviously have more than one force acting on them – they have both their weight pulling them downward and the tension force due to the string pulling them upward. These two forces are known as applied forces and their sum is the net force. Remember that care needs to be taken when adding forces because they have both a magnitude and a direction. The motion of the two masses is not dependent on the applied forces but on the net force. When the two applied forces have the same magnitude but point in opposite directions, the masses do not move regardless if they are at the same height or different heights. When a small additional mass was added, everything accelerated but at a fraction of the acceleration due to gravity.

In all three videos we see a common theme: a net force changes motion. It changes velocity by producing acceleration. That acceleration can be a change in speed, like we saw in the first activity, or it can be a change in the object's direction of motion as will be seen in one of the application activities. The second video also shows us something important. The net force only causes a change in motion, while it actually acts on the object. Once the net force is removed, the object will move at constant velocity, consistent with Newton's first law.

Section 3: Application

App1: Hammer and Feather Experiment on the Moon

(Key Idea: Both objects feel the same acceleration but the hammer has more mass and thus feels more force.)

Directions: Use the video clip to answer the questions below.

1. The astronaut has a hammer in one hand and a feather in the other. Predict what will happen when he lets go. Clearly explain your prediction in the previous question.

2. Now watch the video. Was your prediction correct?

3. Which object has the greater acceleration, greater mass (explain how you know)?

4. Which object feels the greater net force (explain how you know)?

5. If the net force on the hammer is 3.2N and its mass is 2kg, what is the acceleration due to gravity near the surface of the moon? What force would a 0.1 kg feather feel?

App2: Softball Hitter

(Key Idea: The net force is due to the bat, and is equal to the mass of the ball multiplied by the acceleration that it feels. The velocity changes sign as the ball turns around so the acceleration is quite large.)

Directions: Use the video clip to answer the questions below.

1. Based upon your knowledge of softball, when do you think the ball will experience an acceleration other than the acceleration due to gravity?

2. What in the video would support your answer to the previous question?

3. Now, watch the video. Estimate the velocity of the ball before and after it is hit. You can use forward/backward for the direction, or another convention as long as it is clear.

	Magnitude (m/s)	Direction
Before it is hit		
After it is hit		

4. Does the ball appear to accelerate before, after or during the hit?

5. Careful observation will allow you to find the frame just prior to the bat and ball making contact and just after. The precise moment of contact cannot be seen. This sets an upper limit on the amount of time the ball can spend in contact with the bat. Use your estimates of the ball's velocities to estimate the force exerted on the ball by the bat. A softball has a mass of 0.2kg.

App3: Uniform Circular Motion

(Key Idea: A mass moving in a circle is accelerating, because its velocity vector is changing direction. Therefore it must feel a net force, which points inward towards the center of the circle. If you remove the force, the mass will move off in a straight line, consistent with Newton's first law.)

Directions: DO NOT PLAY THE VIDEO CLIPS UNTIL INSTRUCTED.

1. Watch the first video. Do you think the object is accelerating while moving in a circle (explain why)?

2. Does it feel a non-zero net force (if so what provides it)?

3. Now, watch the second video. Is the object accelerating while moving in a circle? What does this tell you about your response to question 1?

4. Explain why the observed motion occurs using Newton's 2nd law.

5. Describe the directions which the velocity, acceleration and net force vectors point for the puck in the first video. Explain why you think each vector points in this direction.
6. observe in the third video.

Lesson 3: Newton's Third Law

Section 1: Exploration

Ex1: Two Trains Crash

(Key Idea: The two cars feel equal and opposite forces because they have roughly equal masses and their accelerations are equal in magnitude and opposite in direction.)

Directions: In the video you'll see two trains of equal mass collide with each other. Use the video to answer the questions below.

1. When the two trains collide does either train feel a non-zero net force?
    a) No, neither train feels a non-zero net force
    b) Yes, one train applies a force and the other train feels that force
    c) Yes, both trains feel a non-zero net force when they collide
    d) You can't tell just from watching the video clip

2. How do you know whether each train feels a force? Explain your reasoning clearly.

3. The trains are moving at 90mph. This is approximately 40m/s. If the two trains each have a mass very near to 15,000kg, estimate the magnitude of the net force felt by each car if the duration of the crash is 0.25s.

Ex2: Ice Skater Propulsion

(Key Idea: The bowling ball and ice skater feel equal and opposite forces. The ice skater is much more massive than the bowling ball, and accelerates much less. The bowling ball is much less massive and accelerates much more.)

Directions: Use the video clip to answer the questions below.

1. Does the ice skater exert a force on the bowling ball, in the horizontal direction, while she's throwing it (explain how you know)?

2. Does the ice skater feel a force in the horizontal direction as a result of her actions (explain how you know)?

3. While the skater is throwing the bowling ball, which object experiences a greater acceleration (explain how you know)?

4. A skater has a mass around 60kg and the bowling ball has a mass of around 7kg. The skater looks to be moving at around 1m/s after she throws the ball. The bowling ball appears to be moving at around 8 or 9 m/s. If it takes the skater about 0.09s to throw the ball, you can estimate the net force felt by the ball and the net force felt by the skater.
    a. Force on the skater (magnitude and direction)
    b. Force on the bowling ball (magnitude and direction)

5. How do the two forces compare?

Ex3: Cart Collisions

(Key Idea: We can calculate that the two carts feel equal and opposite forces when they collide.)

Directions: Use the video clips to answer the questions below.

1. Enter the speeds before and after the collision.
    a. Left cart, speed before.
    b. Right cart, speed before.
    c. Left cart, speed after.
    d. Left cart, speed after.

2. Based on your calculated speed values how would you expect the magnitude of the acceleration of the rightmost cart to compare to the magnitude of the acceleration of the leftmost cart during the collision?
    a) The leftmost cart feels a greater acceleration
    b) The rightmost cart feels a greater acceleration
    c) The two carts feel accelerations that are equal in magnitude
    d) Neither cart accelerates

3. The two carts interact for about 0.2s. Based on this interaction time, estimate the magnitude of the acceleration each cart experiences
    a. Left cart acceleration.

    b. Right cart acceleration.

4. The two carts each have a mass of 0.5kg. Estimate the magnitude of the net force felt on each one using Newton's second law.
    a. Left cart net force.
    b. Right cart net force.

5. How does the magnitude and direction of the net force on the left cart compare to the magnitude of the net force on the right cart?

6. Watch the second video clip. In this video the mass of the right cart has been increased twice. In these two collisions, which cart accelerates more, the one on the right or the one on the left? How do the forces they feel compare?

Section 2: Concept Introduction

(Spoken by SI or read by student)

Let's look at what can be learned from the exploration activities you've just completed.

In the first activity you looked at a video of two trains crashing into each other. It is easy to see that each train exerts a net force on the other. Because their masses are very nearly equal and their accelerations are very nearly equal in magnitude, but opposite in direction, those two net forces are very nearly equal in magnitude and opposite in direction.

This is what Newton's third law tells us: when an object exerts a force on something it feels an equal and opposite force in return. We looked at two equal masses moving with equal initial speeds because that is conceptually easier, but it doesn't have to be that way for Newton's third law to be true, as you saw in the second and third activities.

In the second activity, the skater applies a net force to the bowling ball causing it to accelerate forward, but as a result she can clearly be seen to accelerate backwards. The bowling ball is much less massive than the skater and as a result it feels much more acceleration, consistent with Newton's second law. This video supports that Newton's 3rd law holds true even when the masses are not equal.

In the third activity we look at Newton's third law in a more controlled setting: carts moving on a relatively low-friction track. In the three cases we examined, we have a moving cart colliding with an initially stationary cart. We see that when the carts are equal in mass the moving cart stops and the stationary cart on the right starts moving with nearly the same speed that the cart on the left initially had. This is because they both feel a force as a result of the collision. Since they interact for the same amount of time their accelerations are the same in magnitude, and we therefore conclude that the forces that they feel are equal in magnitude. They are clearly opposite in direction since one cart slows down and the other speeds up. We note that as we increase the mass of the stationary cart we see a more dramatic change in the speed of the moving cart and a less dramatic change in the speed of the stationary cart. If we think conceptually about Newton's second law, the net force is equal to the mass of the object multiplied by the acceleration it experiences; we can see that the behavior we observe suggests that the forces felt by the two carts are still equal and opposite despite the unequal masses. This again suggests Newton's third law is true.

A subtlety that is often overlooked but is easily seen in the lessons is that the forces that form the essence of Newton’s third law, the third law force pairs, act on two separate objects. In the first and third lessons, one force acted on each train or cart. In the second lesson, one force was on the bowling ball while the other was on the skater. If you ever run into two forces that are equal and opposite acting on a single object, they are equal and opposite for reasons other than Newton’s third law.

Another important idea to note is that in all of the cases we've considered in the lessons the Newton's third law force pairs were pretty easy to observe because of changes in motion. We chose these examples for that reason, but Newton's third law is still valid, and often at work, in situations where there are no changes in motion, or no motion at all. Consider a book sitting on a table it feels a weight force down and a support force, or normal force up from the table. Since it feels a force from the table Newton's third law tells us it must also exert a force on the table, which it does (replace the table with your hand to convince yourself), but that force is canceled out by other forces such that the net force on both the table and the book are zero. Don't think that just because everything is at rest there are no forces at play.

Section 3: Application

App 1: Propeller Car

(Key Idea: An object can only accelerate when it feels a net external force. If the force comes from within, then Newton's third law precludes the acceleration of the object.)

Directions: Watch the first video clip, but do not watch the second until instructed.

1. In the first video the cart accelerates when the fan is turned on. Why does this happen? Consider each of the following questions in your explanation: Does the fan exert a force on the cart? Is Newton's third law responsible for this behavior? If so, how?

2. Look at the first picture. Will the cart move when the fan is turned on (why or why not?)?

3. Look at the second picture. Will the car move when the fan is turned on (Why or why not?)?

4. Watch the second video. Were your two predictions correct?

5. Can you explain why the cart moves in the one case, and not in the other? How does this relate to Newton's third law?

App 2: A Car Hitting the Wall

(Key Idea: The wall does feel a force equal in magnitude and opposite in direction to that felt by the car, but the wall is connected to the floor, is connected to the earth, so there is no real change in the motion of the wall.)

Directions: Use the video of the car crashing into the wall to answer the questions below.

1. In the Exploration we clearly observed that when two cars collide they each feel a force due to the other. Newton's third law says that when an object exerts a force on another object it feels an equal and opposite force exerted by the first object. In the two-object collision this is very clear. What about the case when the car collides with the wall. Select the choice that best describes that situation.

a) Newton's third law applies only for free bodies so it doesn't apply to the wall.

b) The wall repels the force and applies it back on the car doubling the effective force on the car.

c) The wall is attached to the floor so although it does feel an equal and opposite force that force is canceled by other forces and the acceleration is essentially zero.

d) The wall's mass is not well defined so the force on it is also not well defined.

2. Explain your reasoning in answering the previous question.

App 3: Two Balls Hitting the Floor

(Key Idea: The bouncy ball has nearly twice the change in velocity, so it feels nearly twice the acceleration and also twice the net force. Since the bouncy ball feels more force, it must also exert more force. Therefore you would want the less bouncy ball dropped on you.)

Directions: Use the video to answer the questions below. The two balls in the video have nearly the same mass.

1. The two balls collide with the floor at nearly the same time. At that time which ball feels a greater force?

   a) The ball that bounces

   b) The ball that doesn't bounce

   c) Both balls feel the same force

   d) You can't tell just by watching the video

2. Explain your reasoning in answering the previous question. How can you tell whether or not the ball's feel different forces?

3. Would you rather have a 5kg ball made of the bouncy or non-bouncy material dropped on your chest? Why?