METACOGNITIVE TUTORING
FOR INQUIRY-DRIVEN MODELING

A Dissertation

Presented to

The Academic Faculty

by

David A. Joyner

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy in Human-Centered Computing in the

School of Interactive Computing, College of Computing

Georgia Institute of Technology

May 2015

METACOGNITIVE TUTORING
FOR INQUIRY-DRIVEN MODELING

Approved by:

Dr. Ashok Goel, Advisor
School of Interactive Computing
Georgia Institute of Technology

Dr. Gautam Biswas
Department of Electrical Engineering and Computer Science
Vanderbilt University

Dr. Ellen Do
School of Interactive Computing
Georgia Institute of Technology

Dr. James Foley
School of Interactive Computing
Georgia Institute of Technology

Dr. Mark Guzdial
School of Interactive Computing
Georgia Institute of Technology

Date Approved: December 18, 2014

Science is more than a body of knowledge. It is a way of thinking;
a way of skeptically interrogating the universe with a fine
understanding of human fallibility.
– Carl Sagan

ACKNOWLEDGEMENTS

Several individuals have had significant influence over the work completed for this dissertation, and for that, I am very grateful.

My wife, Caitlin Joyner, for her love and support.

My adviser, Ashok Goel, for his academic, financial, and personal support throughout my graduate school career.

To the rest of my committee members, Gautam Biswas, Ellen Do, Jim Foley, and Mark Guzdial, for their support and feedback throughout this process, as well as throughout my research career.

Several teachers for their work in running the studies outlined in this dissertation: Kristina Strickland, the teacher for the camp we ran in the summer of 2012 and an adviser to the new teachers using these tools in the spring of 2013; Lara Catall and Vickie Bates, the teachers for the intervention in the spring of 2013; and Angela Gula, the primary coordinator of technology during the intervention.

Georgia Tech's CEISMC, especially Andrew Burch and Chris Thompson, for their work in organizing the summer camp; Mary Moerlins and Helen Ellis, from the Piedmont Park Conservancy and the Georgia Aquarium respectively, for lending their facilities for the 2012 summer camp; Dr. Jeannette Yen, Dr. Marc Weissburg, and Dr. Megan Duffy of Georgia Tech, for also lending their time and energy to the 2012 summer camp.

Dr. Julie Linsey and Dr. Gretchen Rivera, for their consultation and help setting up and conducting the statistical analyses for this dissertation.

Jennifer Kelley, Uzo Okafor, Albert Li, and Sunira Moses, for providing their faces for the metacognitive tutoring system MILA–T.

Several undergraduate and Master's students, for their contributions to the code development of the project, the administration of the summer camp, and the analysis of

data: Nicolas Papin, Cherish Weiler, Suman Manjunath, Rochelle Lobo, Lianghao Chen, Keith Frazer, and Evangelia Spiliopoulou.

The current and former members of the Design & Intelligence Lab, including Swaroop Vattam, Maithilee Kunda, Keith McGreggor, Bryan Wiltgen, Julia Svoboda, David Majerich, Inbal Flash-Gvili, and Spencer Rugaber, for their input and discussions over the past five years.

To my current employer, Udacity, for their patience and understanding as I completed this dissertation during my first year of working with the company, and especially to Jennie Kim, Jason Barros, and Kathleen Mullaney for their support during this year.

You, for reading this.

TABLE OF CONTENTS

LIST OF FIGURES

SUMMARY

Inquiry-based learning approaches have been advocated by researchers in science education for years. Other researchers have also noted the importance of teaching model-based reasoning from an early age. In education more generally, many researchers have noted the importance of developing students' metacognitive skills. Inquiry and modeling have been viewed as metacognitive skills in and of themselves in the past. There exist significant challenges in teaching modeling and inquiry, however, some derived from their nature as metacognitive skills and others derived from the general difficulty in providing guided instruction in open-ended exploratory learning contexts. In order to alleviate these difficulties, this dissertation argues that a metacognitive tutoring system that teaches students an authentic process of inquiry-driven modeling within an exploratory learning activity can address these challenges.

Toward this end, this dissertation first presents a model of the target process of inquiry-driven modeling. It then presents the **M**odeling and **I**nquiry **L**earning **A**pplication (MILA), an exploratory learning environment for facilitating authentic inquiry-driven modeling of an ecological phenomenon. Within this exploratory learning environment and in service of this process of inquiry-driven modeling, this dissertation presents the design and implementation of MILA–Tutoring (MILA–T), a metacognitive tutoring system comprised of five individual artificially intelligent agents for tutoring inquiry-driven modeling.

Given this model of the target process, this exploratory learning environment, and this metacognitive tutoring system, this dissertation presents the design of an intervention to test the effectiveness of MILA–T in teaching the target process. It presents a controlled experiment wherein some teams of students interact with MILA–T and others do not, and it describes the data that were gathered, processes, and analyzed in service of this experiment. Gathering and analysis take place at both the level of the individual student

and at the level of teams of students investigating an ecological phenomenon together, examining the impact of the tutoring system both on individual attitudes and understanding and on the process and products of inquiry-driven modeling within teams.

After describing the structure of and the data gathered through the experiment, this dissertation then describes the results at both the individual and the team level. These chapters examine whether or not interaction with MILA–T helped students compared to interaction with the identical unit and learning environment without MILA–T. For individuals, it first looks at the extent to which interaction with MILA–T impacted individual students' declarative understanding of inquiry-driven modeling. It then looks at the extent to which interaction with MILA–T impacted students' attitudes towards science, scientific inquiry, and careers in science. It also provides some information about the perceptions individual students had about MILA and MILA–T as a whole.

For teams, this dissertation first looks at summaries of the process of inquiry-driven modeling in which teams engage, comparing the processes of teams using MILA–T to the processes of teams interacting without MILA–T. It then looks at the explanations that teams generate during this experiment, comparing the strength and complexity of the explanations generated by teams using MILA–T with those teams generating their explanations without MILA–T. These analyses are repeated for a second project in which no teams interacted with MILA–T in order to check whether any differences diagnosed in the previous analysis were due solely to the presence of extra feedback or due to actual learning of the inquiry-driven modeling process.

This dissertation then brings together the results of these analyses to make an ultimate claim regarding the effectiveness of MILA–T in teaching students inquiry-driven modeling compared to interaction with an identical unit and exploratory learning environment without MILA–T. The effects of MILA–T on metacognition are examined with regard to declarative, dispositional, and procedural knowledge of the target skill; the effect of MILA–T on the explanations generated through the inquiry-driven modeling process is examined as well. Given this ultimate claim, this dissertation then discusses three accounts for how the claimed differences between groups of students actually arose, emphasizing the interaction between individuals, teams, and the classroom as a whole.

Finally, this dissertation discusses the broader contributions of this work to multiple research communities, highlighted by artificial intelligence in education and learning sciences & technology. It then describes a number of future studies that would further corroborate the claims presented herein, as well as studies that would further differentiate between the multiple accounts of the observed differences. Finally, it describes recent applications of the principles featured in the design of this metacognitive tutoring system to online education, as well as new collaborations spawned by this work.

Figure 1, below, presents a pictorial summary of this dissertation. Each column corresponds to a chapter of the dissertation, and each box corresponds to a major section within that chapter. This diagram will be revisited throughout the dissertation to mark progress and emphasize the position of a particular section in the context of the dissertation as a whole.

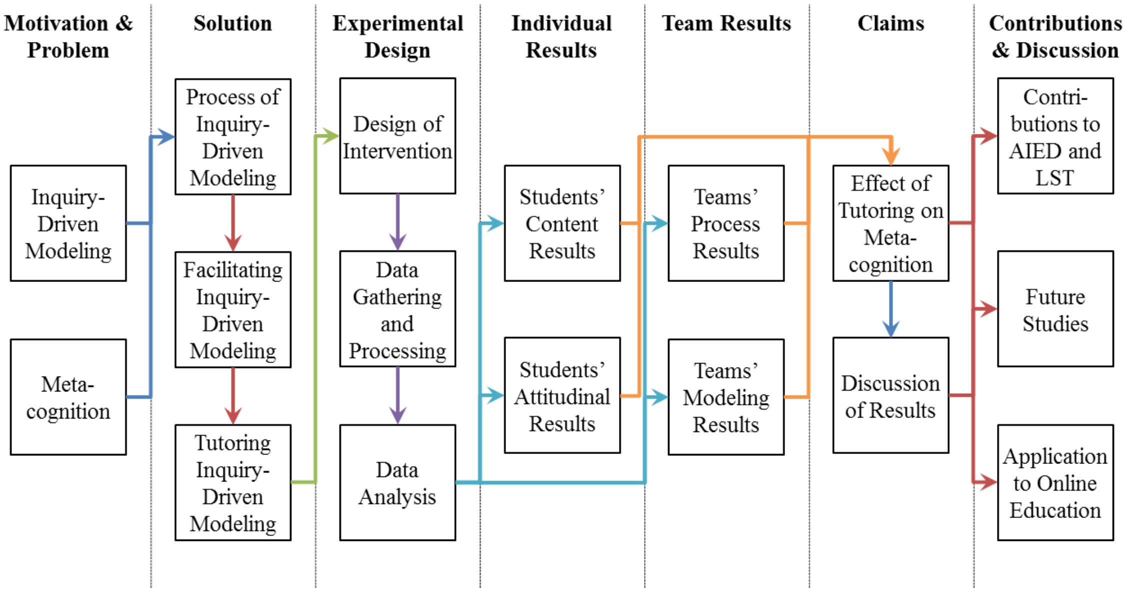

Motivation & Problem
Solution
Experimental Design
Individual Results
Team Results
Claims
Contributions & Discussion
Inquiry-Driven Modeling
Meta-cognition
Process of Inquiry-Driven Modeling
Facilitating Inquiry-Driven Modeling
Tutoring Inquiry-Driven Modeling
Design of Intervention
Data Gathering and Processing
Data Analysis
Students' Content Results
Students' Attitudinal Results
Teams' Process Results
Teams' Modeling Results
Effect of Tutoring on Meta-cognition
Discussion of Results
Contri-butions to AIED and LST
Future Studies
Application to Online Education

Figure 1: A pictorial summary of the contents of this dissertation. Columns across the top represent chapters, while boxes represent major sections. Arrows represent conceptual dependencies: each arrow suggests the target box uses information from the source box. This summary will be used throughout the dissertation to mark progress.

CHAPTER 1

INTRODUCTION & MOTIVATION

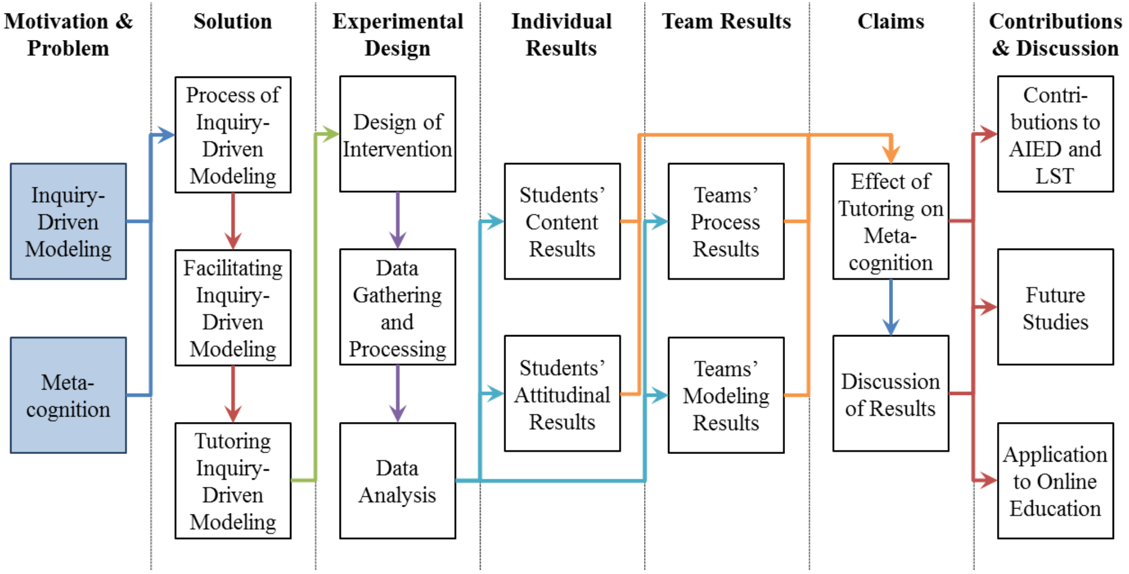

Motivation & Problem
Solution
Experimental Design
Individual Results
Team Results
Claims
Contributions & Discussion
Inquiry-Driven Modeling
Meta-cognition
Process of Inquiry-Driven Modeling
Facilitating Inquiry-Driven Modeling
Tutoring Inquiry-Driven Modeling
Design of Intervention
Data Gathering and Processing
Data Analysis
Students' Content Results
Students' Attitudinal Results
Teams' Process Results
Teams' Modeling Results
Effect of Tutoring on Meta-cognition
Discussion of Results
Contri-butions to AIED and LST
Future Studies
Application to Online Education

Figure 2: The first chapter of this dissertation will cover the motivating problem of this dissertation: teaching students the metacognitive process of inquiry-driven modeling.

This dissertation presents the motivation, design, implementation, deployment, results, and analysis of an initiative to teach the metacognitive skill of inquiry-driven scientific modeling to seventh grade Life Science students in a public Title I school in the state of Georgia. Chapter 1 presents the motivations for the research and the research's problems, questions, and hypotheses. Chapter 2 describes the design and rationale behind an intelligent tutoring system written to help instruct students in this skill. Chapter 3 discusses the intervention itself and its experimental design, logistics, data collection, and objectives. Chapter 4 presents the results of this intervention on individual students' understanding of and disposition toward inquiry-driven modeling. Chapter 5 presents analysis of teams' process of and results from an inquiry-driven modeling exercise. Chapter 6 presents the claims derived from those results and analyses, and Chapter 7 presents the contributions of this work to several communities and a discussion of future applications of this work.

Inquiry-Driven Modeling

In the philosophy of science and science education communities, significant research has been devoted to understanding how scientists make sense of natural phenomena. This field of research draws its motivation from a variety of places, such as the desire to better understand how scientists conduct science and the desire to build better tools to facilitate scientific inquiry and discovery. A major component of this research, however, is educational; recent reforms to science education have placed a much greater emphasis on involving students in authentic scientific practices from an early age, both to increase their interest in science-related careers and to increase their qualifications to pursue such careers in the future (National Research Council 1996; Edelson 1997; Schweingruber, Duschl, & Shouse 2007; Lajoie et al. 2011). Past research has already shown promising results for involving students in authentic processes of science from an early age, such as deeper understanding of complex phenomena after participating in a model-based reasoning exercise (e.g. Goel et al 2013; Lehrer & Schauble 2004).

In practice, scientific modeling is a complex suite of multiple different skills and knowledge structures intended to help scientists make better sense of the systems and phenomena that they analyze. Models play a key role in the process of scientific inquiry; they provide a simplified abstraction of reality that scientists can reason over to generate explanations of their observations, enabling the organization, evaluation, and expansion of current understanding (Darden 1998; Nersessian 1992, 1999, 2008; Griffith, Nersessian, & Goel 2000; Davies, Goel, & Nersessian 2005; Clement 2008). There is no one practice of scientific modeling, but rather the specific processes and strategies of modeling are driven by a diverse set of goals and priorities. (Clement 2008; Nersessian 2008; Svoboda & Passmore 2013). This dissertation chooses to focus on one type of scientific modeling, dubbed here "inquiry-driven modeling". The objectives of inquiry-driven modeling are the construction, manipulation, and improvement of scientific knowledge, both for oneself and for the community as a whole. In inquiry-driven modeling, a scientist constructs a model to make sense of some observed data or phenomenon. The model, then, guides the continued inquiry process, making predictions that can be verified by further data-gathering or emphasizing areas where the mechanism of the model is poorly understood. This high-level notion of the process of inquiry and modeling is compatible with processes observed in the actual course of science (Nersessian 1999). Importantly, this abstract notion is agnostic to the specific types of models used in the inquiry-driven modeling process; the models could be descriptive, prescriptive, conceptual, simulative, causal, dynamic, or of another kind entirely. While the high-level process is agnostic to the type of model used, this dissertation will focus on models that are explanatory and mechanistic (Griffith, Nersessian, & Goel 2000; Goel et al. 2013).

Significant research has identified the value of involving students in authentic scientific processes, such as inquiry-driven modeling, in early science education (Edelson 1997; Gilbert, Boulter, & Elmer 2000; Lee & Butler 2003; Gilbert 2004; Braund & Reiss 2006; Bencze & Hodson 2009). However, such involvement is difficult to implement for several reasons, such as the need to introduce argumentation, negotiation, and justification into the classroom process (Berland & Reiser 2009). As the following section will demonstrate, many of the difficulties with teaching inquiry-driven modeling are derived from its nature as a metacognitive skill and the difficulties in instructing metacognitive skills in general. In order to understand these difficulties and connect them with prior solutions to similar challenges, it is important first to understand the conceptualization of inquiry-driven modeling as a metacognitive skill.

The Metacognition of Inquiry-Driven Modeling

Inquiry-driven modeling can be viewed as a metacognitive skill for three reasons. First and foremost, inquiry-driven modeling meets the definition of a metacognitive skill.

Second, inquiry-driven modeling can be viewed as an instance of other commonly-studied metacognitive skills, most notably self-regulated learning. Third, the facets of inquiry-driven modeling have been discussed as metacognitive skills in past research on inquiry and modeling in education. In the section "Metacognition" on page 8, metacognition will be discussed more generally; the aim of this section is to establish inquiry-driven modeling as a metacognitive skill.

First, inquiry-driven modeling meets the common definitions of metacognition. Brown (1987) defines metacognitive skills as those that address self-assessment of one's own knowledge and regulation of the ways in which to increase or complete that knowledge. Weinert (1987) similarly defines metacognition as "cognition about cognition; that is, it refers to second-order cognitions: thoughts about thoughts, knowledge about knowledge or reflections about actions". As stated previously, the objective of inquiry-driven modeling is the construction, manipulation, and improvement of scientific knowledge, both for oneself and for the community as a whole. It operates both on content knowledge within a given domain (for example, discerning the cause and effect relationships working in an ecological system) as well as on the principles of investigating that domain themselves, such as assessing the reliability of different data-gathering methods or balancing seemingly contradictory observations. Thus, inquiry-driven modeling is a metacognitive skill because knowledge, understanding, and cognition are the targets of the skill.

Secondly, in many ways inquiry-driven modeling can be viewed as an instance of the broader self-regulated learning process, instantiated in a particular domain with an additional set of rules. Self-regulated learning more broadly is a suite of metacognitive skills involving planning, monitoring, and evaluating one's own progress through understanding some new content. Significant research has been devoted to understanding the nature of self-regulated learning and how to develop it in students (e.g. Palinscar & Brown 1984; Butler & Winne 1995; Dweck 2000; Azevedo et al. 2011). Planning, monitoring, and evaluating one's own progress, however, are fundamental to the inquiry-driven modeling process as well. The model that scientists construct serves as a tool to plan further data- and evidence-gathering activities, and the results of those activities are then evaluated against the predictions of the model to check for deviations. Moreover, like self-regulated learning, inquiry-driven modeling is concerned with the processes by which a learner (in this case, a scientist) evaluates their own understanding. From the scientific community, additional rules regarding how to evaluate explanations apply, but the processes of self-evaluation and self-regulation apply within inquiry-driven modeling as well.

Third, in much of the past research on inquiry and modeling (especially in education), these skills have commonly been considered metacognitive in nature. White & Frederiksen (1998) explored attention to metacognition as a way of bringing inquiry and modeling, recognized separately as desirable teaching approaches, into the classroom to make science accessible to a broader range of students. Metacognitive tutoring approaches have also been used in tutoring inquiry (Roll, Aleven, & Koedinger 2010; Gobert et al. 2012), albeit without the modeling component. Schwarz & White (2005) refer to understanding how to use models in inquiry as "metamodeling" knowledge. Metamodeling knowledge, they write, involves understanding the "nature and purpose of scientific knowledge" (p. 166). They go on to say "without such metamodeling knowledge, students cannot fully understand the nature of science, and their ability to use and develop scientific models will be impeded" (p. 166). In this distinction, Schwarz & White differentiate the ability to construct a model from the understanding of the overall role, purpose, and importance of model construction. Inquiry-driven modeling embeds this metamodeling knowledge into its emphases by arguing that the role of modeling is the synthesis of data and observations and the subsequent prediction of further data to gather. Schwarz & White echo this understanding of modeling as part of the broader inquiry process by defining metamodeling knowledge in part as an understanding that "scientific knowledge is a human construct and … models vary in their ability to approximate, explain, and predict real-world phenomena."

The argument that inquiry-driven modeling is a metacognitive skill is important because in connecting inquiry-driven modeling to metacognition, we begin to see that many of the traditional challenges with inquiry-based approaches to education are challenges because of the broader challenges with instructing metacognition in general. By understanding the nature of inquiry-driven modeling as metacognition, we can apply lessons learned from existing projects on teaching other metacognitive skills in other contexts or domains to the learning of inquiry and modeling. In doing so, we see that the challenges faced in developing inquiry- and model-based education are not unique to this skill, but rather more broadly present in metacognitive learning, and thus there is promise in using some of the same approaches used to address metacognition in the past to develop inquiry-based modeling knowledge. However, prior to examining how emphasizing metacognition can help develop knowledge of inquiry-driven modeling, it is important to first address criticisms of inquiry-based approaches to education.

Addressing Challenges and Criticisms in Inquiry in Education

Although inquiry-based approaches to education have garnered significant support and attention over the past several years, there exist significant criticisms on the effectiveness of such approaches. Kirschner, Sweller, & Clark (2006), for example, suggest that inquiry-based interventions involving minimal guidance do not work because initially, learners do not have a sufficient background in the target skill to provide their own guidance. Klahr & Nigam (2004) directly compared discovery learning, a method for inquiry-based learning (Bruner 1961), with direct instruction and found that students participating in direct instruction activities demonstrated significantly better outcomes than students in discovery-based activities. Mayer (2004) similarly recounts the failure of pure discovery learning in three separate contexts. Given that this dissertation focuses precisely on inquiry- and discovery-based learning strategies, it is important to first describe how this research reacts to the criticisms of this broad approach.

First, the existing criticism of inquiry-based approaches largely criticizes the use of inquiry-based learning in service of separate learning goals. For example, if the learning goal of a particular exercise is to instruct students on how and why fish died in a lake (as will be the example in this research), these critics argue it is significantly more effective to directly instruct students on how and why this occurred than to have them discover it for themselves. This criticism assumes that the learning goal is a correct understanding of the phenomenon, and in many units this assumption is correct. However, the learning goal of this research differs. In this work, correct understanding of some content knowledge is *not* the goal, but rather, the goal is the inquiry-driven modeling skill itself. Inquiry-based approaches have been criticized as being poor means to a separate end, but this research treats the inquiry-driven modeling process as the end in and of itself. The goal here is to teach inquiry-driven modeling rather than to use inquiry-driven modeling to teach some underlying content knowledge. Given the research discussed earlier that identifies understanding the nature and practice of science as valuable learning goals in their own right (Schwarz & White 2005), and given the alignment of state and national standards to this assertion (National Research Council 1996; Edelson 1997; Gilbert, Boulter, & Elmer 2000; Lee & Butler 2003; Gilbert 2004; Braund & Reiss 2006), this learning goal largely avoids this criticism of inquiry-based learning. In fact, given the assertion of Kirschner, Sweller, & Clark that "the advantage of guidance begins to recede only when learners have sufficiently high prior knowledge to provide 'internal' guidance", the goal of this work can be seen as an attempt to address precisely the concern that these critics have raised. If it is true that inquiry-based learning is only effective once students know enough about inquiry to provide internal guidance, then a curriculum such as this one that aims to teach inquiry can serve as the bridge that allows other inquiry-based approaches to find greater success.

However, the criticism applied to interventions that use inquiry as a means to an end could equally be applied if this intervention used inquiry as a means to teach inquiry; in other words, the prior criticisms would be equally valid here if we attempted to allow students to "discover" the value of the inquiry process on their own. This work hypothesizes, however, that given an environment in which students could participate in inquiry-driven modeling, additional instruction on the proper participation in the process is still necessary. While students may demonstrate some rudimentary understanding of inquiry and modeling in the absence of guidance, guided instruction is critical to developing a true understanding of the target process. Thus, this work acknowledges these criticisms and attempts to address them specifically by providing the kind of guided instruction that Kirschner, Sweller, & Clark advocate. In fact, this work can be viewed specifically as an attempt to provide distributed guided instruction to multiple teams engaged in the inquiry process simultaneously.

Metacognition

Establishing inquiry-driven modeling as a metacognitive skill allows the existing research on metacognitive development to be applied to the inquiry domain. Developing metacognition has been identified as a key goal of instructional design (Baker & Brown 1984; Bransford, Brown, & Cocking 2000), and many modern theories of learning and achievement point to the importance of metacognitive skills and mindsets in predicting and determining students' ongoing success (Babbs & Moe 1983; Borkowski, Carr, Rellinger & Pressley 1990; Ganz 1990; Landine & Stewart 1998; Sternberg 1998; Dweck 2006; Isaacson & Fujita 2006). While developing metacognition is important, it also presents unique challenges. Several initiatives have been successful at developing metacognitive skills (e.g. Borkowski, Chan, & Muthukrishna 2000; Schraw, Crippen, & Hartley 2006; Lovett 2008; Vattam et al. 2011A; Goel et al. 2013), but often present their own unique challenges, such as high costs and low scalability.

Challenges in Developing Metacognitive Skills

Metacognitive skills have several unique traits that differentiate them, especially in terms of instruction, from other skills. A strong description of the several of these differences and the challenges they present can be found in Roll et al. 2007. Roll et al. specifically propose four fundamental factors that differentiate metacognitive skills from cognitive skills. First, they tend to be domain independent. Skills such as help-seeking, self-assessment, self-explanation, and content structuring are domain-independent and affect learning in different subjects and contexts; however, learning is typically embedded in some specific subject, meaning that metacognitive skills must be taught in the context of other skills, competing for students' attention. Second, metacognitive skills are difficult to teach explicitly: their nature as operators on content knowledge or cognitive skills means that they typically are taught in conjunction with more recognizable, assessable skills. A third related problem is that these metacognitive skills are often deemphasized in the eyes of students in comparison to the more tangible, assessable cognitive skill (also in Gama 2004); this connects to the points above in that metacognition is typically taught in the context of a cognitive skill or some content knowledge that is more easily identifiable by students as the ultimate learning goal. Fourth, metacognitive skills exist largely "in the head" of the individual, meaning that the same external action might be "correct" for one student and "incorrect" for another, in that the underlying thought process that gave rise to the same behavior might have been different. This introduces a significant problem in assessment in that the visible behaviors on which tutoring systems most often act are merely shadows of the student's underlying skills and abilities. This also presents a significant challenge in researching metacognition as it can often be difficult to ascertain when metacognitive ability has actually been improved; this is discussed more in the section "Challenges in Researching Metacognition" on page 12.

Other challenges exist as well: it can be difficult to specifically describe objective efficacy with the skill (Roll et al. 2007); effective learning involves demonstration and increased participation in the skill rather than simply instruction and lecture (Lave & Wenger 1991; Spiro et al. 1992; Anderson, Reder, & Simon 1996); and it can be difficult to establish a baseline of students' prior performance with metacognitive strategies (Biswas et al. 2013). As described in Roll et al. 2012, many interventions, such as inquiry-based learning environments or project-based learning approaches, focus on facilitating the development and application of metacognitive skills while lacking any explicit attention to teaching those metacognitive skills, due in part to the difficulties in teaching such skills outlined above. These issues resonate strongly with the criticisms given in Kirschner, Sweller, & Clark 2006: providing an environment in which exploratory learning can occur is not the same as teaching students how to actually participate in exploratory learning. Compounding these issues as well are the previously-documented challenges with teaching groups as effectively as individuals (Bloom 1984); even if there exist initiatives that allow students to effectively learn metacognitive skills with a one-on-one instructor (such as the apprenticeship approaches described by Lave & Wenger 1991), extensions of those initiatives to group learning and classrooms remain elusive.

There are parallels to these challenges in the instruction of inquiry-driven modeling, as discussed above. Although inquiry-driven modeling is typically taught within a specific content domain (such as ecology, chemistry, or physics), it spans across every scientific discipline, making it more domain-neutral than other learning goals present in early science education. Explicit instruction is often challenging because the components of inquiry-driven modeling are within the scientist's head rather than explicitly out in the world, and thus, demonstration, mentorship, and apprenticeship are all valuable despite their challenges. Similarly, students are often entirely consumed with arriving at the right answer, but inquiry-driven modeling emphasizes that answers must be defended and justified, not simply labeled as correct or incorrect; thus, students are not naturally predisposed to value, or even think in terms of, the metacognitive skill. These parallels help demonstrate two takeaways: first, they further emphasize that inquiry-driven modeling is a metacognitive skill as seen in the shared instructional difficulties, and second, that the existing attempts to solve these problems in other metacognition-focused interventions may apply to improving the instruction of inquiry-driven modeling as well.

Intelligent Tutoring Systems as a Potential Solution

Although there remain difficulties, intelligent tutoring systems do provide solutions to many of the challenges inherent to teaching these metacognitive skills. Intelligent tutoring systems, generally, have been suggested as a solution to the challenge of extending effective individual instruction to group and classroom learning (Corbett 2011). In addressing those challenges, software provides the opportunity to track students' progress persistently and individually rather than relying on intermittent assignments and summative assessments. Similarly, feedback can be provided individually during the activity itself. In this way, students can receive just-in-time feedback, which has been shown to improve learning outcomes (Anderson, Corbett, Koedinger, & Pelletier 1995).

Intelligent tutoring systems for cognitive skills have been successful for many years (e.g. Ritter, Anderson, Koedinger, & Corbett 2007; Matsuda & VanLehn 2005; Crosby & Iding 1997; Aleven, McLaren, Sewall, & Koedinger 2009); in recent years, these approaches have begun to be applied to metacognitive skills. At present there exist several successful initiatives in metacognitive tutoring. HelpTutor teaches help-seeking in the context of the established Cognitive Tutoring project (Aleven, McLaren, Roll, & Koedinger 2006; Roll et al. 2007), where research is also done on students' tendencies to "game" the tutor to get the most information out as quickly as possible (Baker, Corbett, Roll, & Koedinger 2008). MetaTutor teaches self-regulated learning in the context of an interactive hypermedia on biological systems (Azevedo et al. 2009) and has made significant strides in designing ways of detecting students' mental models of metacognitive skills (Lintean, Rus & Azevedo 2012) and reacting by modifying the available content (Azevedo et al. 2011). Betty's Brain turns the relationship on its head by putting the student in the position of tutoring a software agent, focusing in part on metacognitive reasoning (Biswas et al. 2005; Kinnebrew, Biswas, Sulcer, & Taylor 2012). Other initiatives focus on other metacognitive skills, such as learning from errors (Mathan & Koedinger 2005), self-explanation (Aleven & Koedinger 2002), and self-assessment (Roll, Aleven, McLaren, & Koedinger 2011).

Challenges in Researching Metacognition

Interestingly, many of the challenges inherent to teaching metacognition are also challenges to researching metacognition. Metacognition is largely invisible, existing only in the head of the learner, and although observable acts can be used to estimate students' metacognitive processes, there will always exist a layer of interpretation between the observable actions and the perceived metacognitive ability. One way in which this has been resolved in the past is through structuring students to make their metacognitive processes explicit; MetaTutor (Azevedo et al. 2009) accomplishes this by supplying students with tutors that specifically demonstrate elements of the metacognitive process that students ought to learn to adopt; this approach resonates with the apprenticeship learning approach advocated by other research (e.g. Lave & Wenger 1991). In doing so, students are asked to make some of their metacognitive processes, like planning and monitoring, explicit. This serves two functions: the process of making them explicit helps students value and execute the skills, and having explicitly-stated plans and monitoring habits provides information on which the tutors of MetaTutor can act.

The process of creating explicit plans and monitoring behaviors introduces two mechanics, however. First, the simple process of making these explicit creates a learning opportunity. This opportunity is valuable, but it creates difficulty identifying the exact role of tutoring in supporting the development of metacognitive processes. An ideal environment for facilitating metacognition would likely involve this component, but for research purposes this introduces a separate mechanism for student improvement that makes measuring the effect of the tutoring system itself more difficult. Secondly and more importantly, providing a way in which students can make their plans and monitoring behaviors explicit constrains the overall learning environment to those plans and behaviors that the tutoring system can understand. Although this is often valuable to help students learn the desired skill, it also opposes some of the appeal and value of an open-ended exploratory learning environment for inquiry-based learning.

The research on metacognition presented in this dissertation explicitly attempts to examine these areas that are largely untouched by other tutoring systems. In doing so, the tutoring system developed here will be at an active disadvantage in some ways because the information available to it is less explicit; the tutors developed for this research will have to infer far more about students' metacognitive processes than other metacognitive tutoring systems. Although the tutors will actively instruct students to engage in critical metacognitive processes like planning and monitoring, they will not have an explicit means by which to identify students' metacognition; instead, they will have to infer metacognition from observable actions. However, by allowing students to proceed without making their metacognitive processes explicit, this research will more authentically expose students to the open-ended and exploratory nature of authentic scientific modeling and inquiry. In this way, this research favors authenticity over measurability; although metacognition will be difficult to measure in this work, the process in which students engage is hypothesized to be more similar to the process in which scientists engage in the real world.

Research Agenda

The goal of this research is to unify these two broad research communities – intelligent tutoring systems and inquiry-driven modeling – in service of one another. First, in leveraging inquiry-based approaches to scientific education, this intervention aims to build on the value and appeal of such inquiry-based approaches to science education while also improving the state of the art in inquiry-driven education. However, this goal is also established with an awareness of the existing difficulties in and criticisms of using inquiry-based approaches in science classrooms (e.g. Kirschner, Sweller, & Clark 2006). These difficulties and obstacles bring in the emphasis of this research on developing a metacognitive tutoring system to facilitate learning the inquiry-driven modeling process. Through a tutoring system embedded in an exploratory learning environment in which inquiry-driven modeling can occur, this research aims to overcome the problems of providing guidance, feedback, and apprenticeship to multiple student teams simultaneously in a classroom context.

Research Problem

As stated above, inquiry-driven modeling, as with any metacognitive skill, has significant challenges to instruction, including the difficulty in providing consistent feedback, the difficulty in creating an environment in which the student can be an increasing participant in the process, and the difficulty with persistently monitoring and modeling students' ability. These challenges are especially notable in traditional classroom environments where one individual (typically the teacher) is responsible for the learning of several students. It has further been noted as well that in disciplines with established standards, such as scientific research, it is important to directly instruct students on proper participation in the discipline rather than relying on them to discover the proper procedures on their own (Shulman & Keisler 1966; Mayer 2004). Although this has most commonly been cited as a challenge in using inquiry-based approaches to teach content knowledge or cognitive ability (Klahr & Nigam 2004), the difficulty remains even when inquiry is the learning goal itself rather than a means to a separate end.

The problem, then, of this research is to design ways in which guided instruction of inquiry-driven modeling can be implemented in classroom learning environments, overcoming the challenges described above. Embedded in this research agenda is the hypothesis that, given an environment in which teams of students can participate in inquiry-driven modeling, a metacognitive tutoring system will be able to provide the guided instruction needed to learn and master the process. This intention to use a metacognitive tutoring system to address the research problem then introduces additional problems. Inquiry-driven modeling is a metacognitive task that takes place largely in the mind of the student or in the conversations amongst teams of students, and thus is simultaneously difficult to observe and difficult to demonstrate. Teams of students are often more focused on the content knowledge or the correct answer than the metacognitive process that gives rise to an answer. Teams of students may also have fundamental metacognitive misconceptions about the nature of inquiry and the process of science which themselves are difficult to discern. Thus, a metacognitive tutoring system for inquiry-driven modeling must be able to observe, model, demonstrate, and instruct a difficult, oftentimes invisible skill that students are not naturally predisposed to emphasize, value, or even recognize. Adding to this difficulty, authentic participation in this skill is almost universally team-based. The emphasis on authenticity in this research shows the need to teach this skill in a team-based activity as well, but gathering information at the team level introduces additional challenges with regard to measuring individual students' performance and understanding. Thus, from the nature of the skill to the context in which it is demonstrated to the emphasis placed on the skill by the students, designing a metacognitive tutoring system for inquiry-driven modeling introduces numerous challenges.

Research Questions

The questions addressed by this dissertation, in turn, examine the effectiveness of a metacognitive tutoring system in teaching inquiry-driven modeling. Compared to a comparable activity without a metacognitive tutoring system, how effective is a metacognitive tutoring system for inquiry-driven modeling at improving students' understanding of the skill? In order to address this question, however, it is necessary to first articulate what it would mean to understand the skill. Prior literature on metacognitive tutoring suggests examining at least three facets of a metacognitive skill: declarative knowledge, procedural knowledge, and dispositional framing (Roll et al. 2007). Mastery of a metacognitive skill, by this categorization, includes being able to describe the skill (declarative knowledge), being able to execute the skill (procedural knowledge), and having the desired attitudes towards the value and appropriate application of the skill (dispositional framing). Thus, these three dimensions must be considered. In addition, the process of inquiry-driven modeling is important in the scientific community because of the quality of the explanations, theories, and models that it generates. Thus, in order to assess the benefit of a metacognitive tutoring system for inquiry-driven modeling, it is also important to consider the effect of the system on the quality of the output that the process creates.

Based on these components of metacognitive ability and the function of the process in the broader scientific community, this articulation presents four distinct targets of analysis to assess the effectiveness of a metacognitive tutoring system: explicit understanding, dispositional framing, procedural execution, and final explanations. However, it is important to note that the last two of these targets have two dimensions to each. First, we can examine whether or not improvements were seen while the metacognitive tutoring system was in action: while receiving feedback from a metacognitive tutoring system, does the process in which students or teams of students engage improve? Do the models that they produce improve? However, this question does not identify whether any improvements seen are simply functions of additional feedback or functions of an internalized improved understanding. To illustrate this with a more accessible example from a different domain and a distributed cognition explanation (Hollan, Hutchins, & Kirsh 2000), a student performing better on Algebra problems when using an Algebra tutor merely suggests that the system comprised of the student and the tutor is better than the student alone. In order to test true improvement, one would examine whether the student's performance improved in the absence of the tutor based on prior interaction with the tutor.

A second question, then, follows: do students or teams of students that previously received feedback from a metacognitive tutoring system show ongoing improved performance even after the tutoring system has been disabled? Are the models and explanations they produce superior even in the absence of tutor feedback on those specific models and explanations? This leads to two dimensions of these hypotheses that will be examined throughout this dissertation, the Learning phase and the Transfer phase. The Learning phase examines whether students or teams of students perform better while receiving the additional feedback (i.e. while Learning the skill), while the Transfer phase examines whether these performance improvements transfer to a new task wherein students are no longer actively receiving that feedback.

This dissertation, then, aims to answer the broad question of how interaction with a metacognitive tutoring system for inquiry-driven modeling affects students' performance across these four different metrics, with special attention paid to changes both during interaction with the metacognitive tutoring system and after the system has been disabled. Table 1, below, provides a more detailed description of the eight research questions implicit in this broad question, its four components, and its two dimensions.

Table 1: Research questions of this dissertation.

Explicit Understanding	
Question #1: To what extent does engagement with a metacognitive tutoring system during participation in an inquiry-driven modeling task improve students' explicit understanding of the inquiry-driven modeling process?	
Dispositional Framing	
Question #2: To what extent does engagement with a metacognitive tutoring system during participation in an inquiry-driven modeling task change students' attitudes toward science, scientific inquiry, and future careers in science-related fields?	
Procedural Execution	
Question #3: To what extent does engagement with a metacognitive tutoring system during participation in an inquiry-driven modeling task alter student teams' successful execution of the desired inquiry-driven modeling process?	
Question #3A: To what extent is students' successful execution of the desired inquiry-driven modeling process altered **during** interaction with the metacognitive tutoring system?	**Question #3B:** To what extent is students' successful execution of the desired inquiry-driven modeling process altered **after** interaction with the metacognitive tutoring system?
Models and Explanations	
Question #4: To what extent does engagement with a metacognitive tutoring system during participation in an inquiry-driven modeling task alter the explanations that teams of students generate?	
Question #4A: To what extent are the explanations students generate altered **during** interaction with the metacognitive tutoring system?	**Question #4B:** To what extent are the explanations students generate altered **after** interaction with the metacognitive tutoring system?

Research Hypotheses

In response to these research questions, this dissertation initially adopts the broad hypotheses that participation in an exercise in inquiry-driven modeling using a metacognitive tutoring system will improve students' understanding of inquiry-driven modeling compared to students who participate in the identical exercise without the presence of such a metacognitive tutoring system. Specifically, this improved understanding will be seen across four dimensions: compared to students who do not have access to the tutoring system, students who do interact with it will have an improved declarative understanding of the process, a more positive dispositional framing of the process, a superior procedural execution of the process, and higher quality results of the process. Moreover, this improved procedural execution and these higher-quality results will be present both while the students are receiving feedback from the tutoring system as well as in a new exercise in which students are no longer receiving feedback. These hypotheses are outlined in further detail in Table 2, below.

This dissertation will begin in Chapter 2 by first describing the design of a system for providing metacognitive tutoring for this inquiry-driven modeling process. In order to do this, it will first provide a desirable model for inquiry-driven modeling grounded in the literature and past research on this project. In order to ground this model and the subsequent design of a metacognitive tutoring system, this dissertation will also briefly describe the design of an exploratory learning environment, MILA (**M**odeling & **I**nquiry **L**earning **A**pplication) in which this inquiry-driven modeling process can be executed and tracked. Based on this model and this exploratory learning environment, this dissertation will then describe the design of a metacognitive tutoring system, MILA–T (MILA-**T**utoring), embedded in MILA that guides and instruct students towards an improved understanding of the inquiry-driven modeling process.

Table 2: Research hypotheses of this dissertation.

Explicit Understanding	
Hypothesis #1: Engagement with a metacognitive tutoring system during participation in an inquiry-driven modeling task will improve students' declarative understanding of inquiry-driven modeling compared to participation in inquiry-driven modeling without a metacognitive tutoring system.	
Dispositional Framing	
Hypothesis #2: Engagement with a metacognitive tutoring system during participation in an inquiry-driven modeling task will improve students' dispositional framing of science, scientific inquiry, and careers in science compared to participation in inquiry-driven modeling without a metacognitive tutoring system.	
Procedural Execution	
Hypothesis #3: Engagement with a metacognitive tutoring system during participation in an inquiry-driven modeling task will improve teams' execution of the desired inquiry-driven modeling process compared to participation in inquiry-driven modeling without a metacognitive tutoring system.	
Hypothesis #3A: This improved execution of the desired inquiry-driven modeling process will take place while the team is receiving feedback from the metacognitive tutoring system.	**Hypothesis #3B:** This improved execution of the desired inquiry-driven modeling process will take place when the team is no longer receiving feedback from the metacognitive tutoring system.
Models and Explanations	
Hypothesis #4: Engagement with a metacognitive tutoring system during participation in an inquiry-driven modeling task will improve the quality of models and explanations that teams generate compared to participation in inquiry-driven modeling without a metacognitive tutoring system.	
Hypothesis #4A: These improvements will be present in modeling activities during which the team is receiving feedback from the metacognitive tutoring system.	**Hypothesis #4B:** These improvements will be present when the team is no longer receiving feedback from the metacognitive tutoring system.

After describing the tools designed for this project, Chapter 3 will describe the experiment used to test whether or not the results from deployment of these tools match the hypothesized results from Table 2, above. In order to test the accuracy of these hypotheses, a controlled study will be used comparing teams of students interacting with the exploratory learning environment *without* the metacognitive tutoring system (the Control group) to teams interacting with the exploratory learning environment *with* the metacognitive tutoring system (the Experimental group). Besides the presence or absence of the tutoring system, all other elements of the activity completed by the two groups of students will be held constant.

Given the experimental design described in Chapter 3, Chapters 4 and 5 will describe the results of this experiment. First, Chapter 4 will establish that the groups of students participating in the intervention were effectively equal between the Control and Experimental groups at the start of the intervention. After establishing that foundation, Chapter 4 will examine the impact of the intervention on individual students. It will first look at how participation in the intervention affected students' understanding of inquiry-driven modeling, focusing on whether or not engagement with MILA–T improved understanding compared to engagement in the intervention without MILA–T. It will then similarly look at how participation in the intervention affected students' dispositions toward science, scientific inquiry, and careers in science, and especially whether engagement with MILA–T affected these dispositions. It will finally look at general data from students on the extent to which they perceive MILA and MILA–T as beneficial to their inquiry-driven modeling process.

Chapter 5 will present similar analyses and results, but at the team level. Chapter 5 will first examine the process of inquiry-driven modeling in which teams of students engaged, using Markov chains to capture the overall process and connect it with the process of inquiry-driven modeling (and errors therein) shown in Chapter 2. It will then compare these chains between Control and Experimental teams, testing for the degree to which each group of teams matched the desirable inquiry-driven modeling process. Chapter 5 will then turn to the explanations that students actually generated, evaluating them across five different metrics as well as coding the justifications teams provide for accuracy and relevance. In particular, it will examine whether teams interacting with MILA–T generate better explanations for the target phenomena than teams that do not have access to MILA–T. Finally, Chapter 5 will also report on the perceptions of the teachers in the classroom on the results of the intervention.

Given the results of the analysis presented in Chapters 4 and 5, Chapter 6 will synthesize the results into a set of claims based on the data and original hypotheses. It will first run through the claims provided by each hypothesis individually, summarizing the analyses behind each and restating the conclusions to each hypothesis. It will then combine the results of those individual analyses into a broader claim about the impact of the intervention on students' metacognitive ability.

Given the claims presented in Chapter 6, Chapter 7 will then discuss the contributions of this work to different communities, emphasizing artificial intelligence for education and learning sciences & technology, but also referencing intelligent tutoring systems, exploratory learning environments, and other communities. Finally, Chapter 7 will discuss three directions for further research building on the results of this study: first, the suggested next steps for examining the new hypotheses derived from this work; second, recent applications of the same motivating principles in a different domain; and third, new collaborations spawned by this work with the Georgia Tech School of Biology, the Georgia State University College of Education, and the Smithsonian Institution.

CHAPTER 2

INTELLIGENT TUTORING SYSTEMS AS A SOLUTION

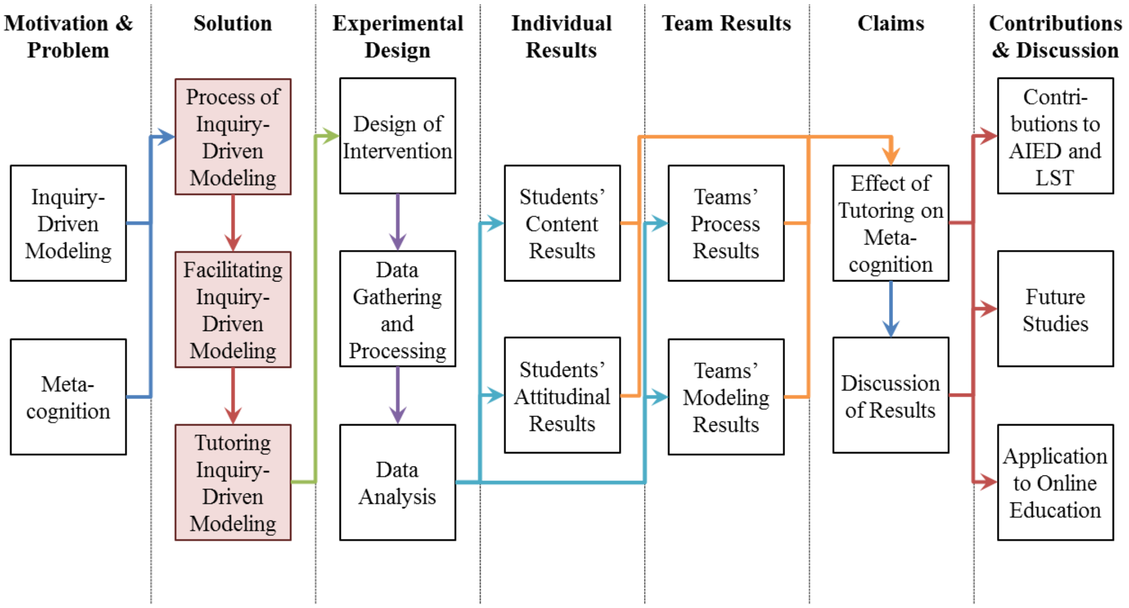

Motivation & Problem
Solution
Experimental Design
Individual Results
Team Results
Claims
Contributions & Discussion
Inquiry-Driven Modeling
Meta-cognition
Process of Inquiry-Driven Modeling
Facilitating Inquiry-Driven Modeling
Tutoring Inquiry-Driven Modeling
Design of Intervention
Data Gathering and Processing
Data Analysis
Students' Content Results
Students' Attitudinal Results
Teams' Process Results
Teams' Modeling Results
Effect of Tutoring on Meta-cognition
Discussion of Results
Contri-butions to AIED and LST
Future Studies
Application to Online Education

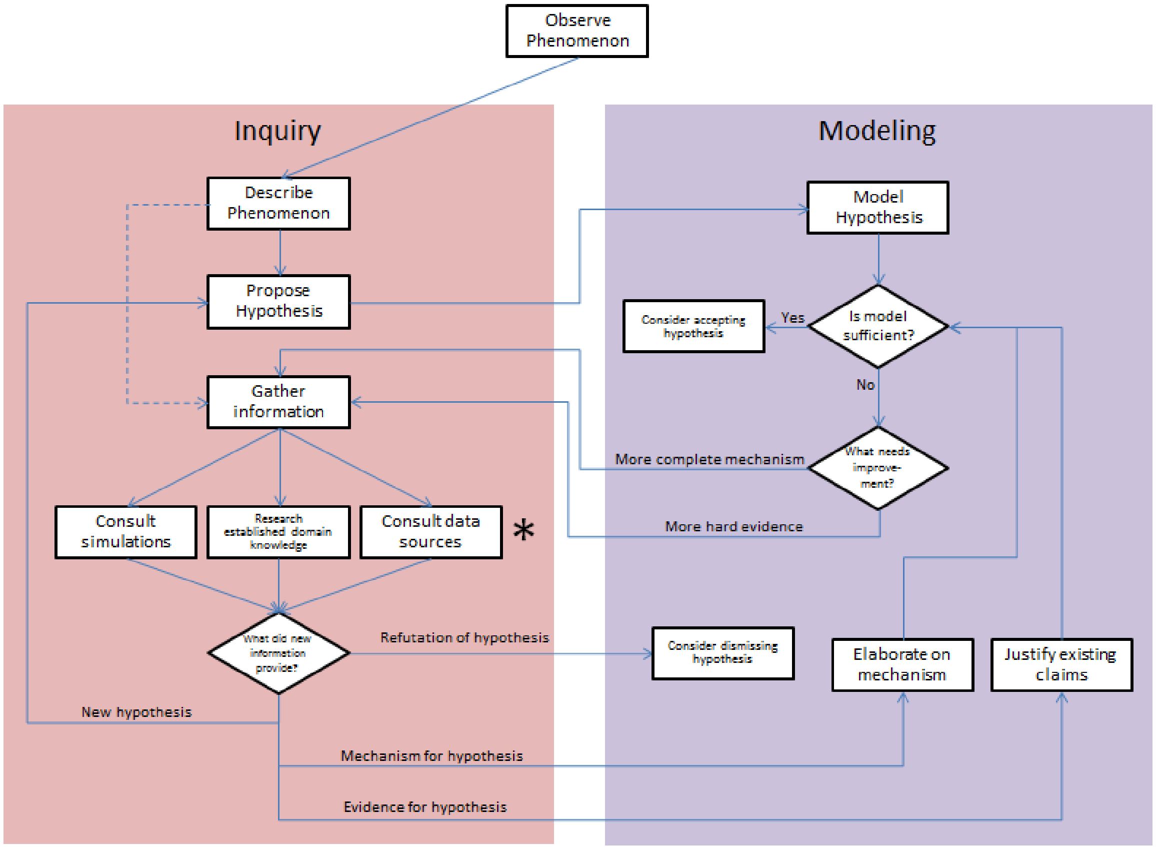
Figure 3: The second chapter of this dissertation covers the solution to the problem posed previously: a formalization of the process of inquiry-driven modeling, an exploratory learning environment that can facilitate inquiry-driven modeling, and most importantly, the design of a metacognitive tutoring system to teach students inquiry-driven modeling.

The objective of the design portion of this project is to create a metacognitive tutoring system that, when situated in an exploratory learning environment that facilitates inquiry-driven modeling, improves students' understanding of that target skill. Prior to developing such an intelligent tutoring system, however, two components are necessary: a model of the skill it strives to teach, and an environment in which it can teach the skill. Toward this end, this chapter will first describe the process and results of formalizing a model of a desirable process. Given the model, it will then describe the design of an exploratory learning environment in which the desired process of inquiry-driven modeling can be executed and monitored. With these two tools in place, this chapter will then move to describe in detail the implementation of the metacognitive tutoring system, first at a high architectural level and subsequently in more detail for each of five independent pedagogical agents.

Process of Inquiry-Driven Modeling

In order to effectively teach students this authentic process of inquiry-driven modeling, it is first necessary to have a model of what the process itself looks like, both to guide the construction of the exploratory learning environment and accompanying tutoring system and to provide a process to communicate to the students explicitly (Roll, Aleven, McLaren, & Koedinger 2007). In the community, there are several general theories regarding what the scientific process actually looks like and what it ought to look like, but presently there does not exist a process model of a desirable execution of inquiry-driven modeling. Thus, this project first starts by developing such a model. This model is developed based on three primary factors: the existing literature on inquiry and modeling, our own prior experience with modeling education across several different exploratory learning environments, and our own observations from a pilot study testing early versions of these tools. This section will first present and explain the model of the process of inquiry-driven modeling in Figure 4; it will then explain the development of this model based on the three sources described above.

In the process model shown in Figure 4, boxes represent tasks (such as 'Observe Phenomenon', 'Describe Phenomenon', and 'Propose Hypothesis), while diamonds represent decisions (such as 'Is model sufficient?' and 'What needs improvement?'). Edges within the model represent transitions between tasks. Annotated edges coming out of decision diamonds represent the answer to the question posed in the diamond; for example, when students face the decision 'Is model sufficient?', they can answer either 'Yes' (and thus enter the 'Consider accepting hypothesis' task) or 'No' (and thus enter the 'What needs improvement?' decision). The dotted line between 'Describe Phenomenon' and 'Gather Information' is an optional pathway advocated by some literature on scientific modeling and inquiry (Nersessian 2008) which suggests scientists gather initial information before ever proposing a hypothesis to explain the phenomenon. In the event that this pathway is taken, the decision 'What did new information provide?' can only have one possible exit ('New hypothesis'), given that the other three exits ('Refutation of hypothesis', 'Mechanism for hypothesis', and 'Evidence for hypothesis') all rely on a hypothesis previously existing. The three (and potentially more) tasks under 'Gather Information' can be considered subtasks or types of the 'Gather Information' task.



* - Numerous other sources of information possible; only a subset is displayed here.

Figure 4: A model of a desirable process of inquiry-driven modeling derived from the literature and experience in the pilot study.

Inquiry-Driven Modeling Walk-Through

To illustrate this model in action, we will walk through an example of a team engaging in the process described above. As described later in this dissertation, in this experiment, teams of students attempt to explain a sudden massive fish kill in Lake Clara Meer. A fish kill is an event wherein a massive number of fish simultaneously die in a localized area (Townsend, Boland, & Wrigley 1992), often due to oxygen depletion (Barica & Mathias 1979). Such an event occurred in a lake near Georgia Tech in 2009

(Wheatley 2009). During the intervention described in Chapter 3, teams of students attempted to construct models of what may have caused this fish kill. The following example uses this phenomenon to explore the model of inquiry-driven modeling given in Figure 4.

To start, the investigator (a student or a scientist) begins the process by observing some phenomenon that they wish to explain; in this case, they observe the fish kill through a video played at the start of the unit. They then come to a clean, delineated description of the phenomenon to be described to focus the modeling and information-gathering processes. In this project, the phenomenon could be described as: "Thousands of fish suddenly died in Lake Clara Meer." The definition of the phenomenon is important because it helps direct students' attention to the salient parts of the phenomenon to be explained; teams with poor phenomenon definitions sometimes became distracted by how the smell of the fish drove visitors away from the park, a phenomenon separate from the one to be described.

After defining the phenomenon to be explored, investigators may either propose a hypothesis or go ahead and begin data gathering. Some models of inquiry (Nersessian 2008) suggest scientists rarely propose hypotheses with no information about the phenomenon to be described, and thus information-gathering always takes place first; others may argue that even describing the phenomenon in the first place is a type of data-gathering task. If the investigator does propose an initial hypothesis before gathering additional information, however, they construct a preliminary model of it, typically involving nothing but the conjecture that their chosen hypothesis does cause the phenomenon. In this example, a team might hypothesize that pollution was responsible for the fish kill. In this case, this model would simply show a causal connection between rising pollution and declining fish populations. This model, however, is insufficient based on model evaluation criteria given to students during the unit, and thus they return to the information-gathering task.

Given an insufficient model, the investigator returns to the information-gathering task, looking for information that will help elaborate on their model. In this example, the team has proposed pollution as the cause, so they attempt to find information on how pollution kills fish by consulting established domain knowledge. During this consultation, however, they find that current models of pollution's effect on aquatic ecosystem suggest that pollution slowly destroys ecosystems over time rather than killing all fish at once. Moreover, pollution destroys all life in a system, not just the fish. This gives them new information to use to evaluate their models, asking themselves 'What did new information provide?' This information refutes their hypothesis, which leads them to dismiss this hypothesis.

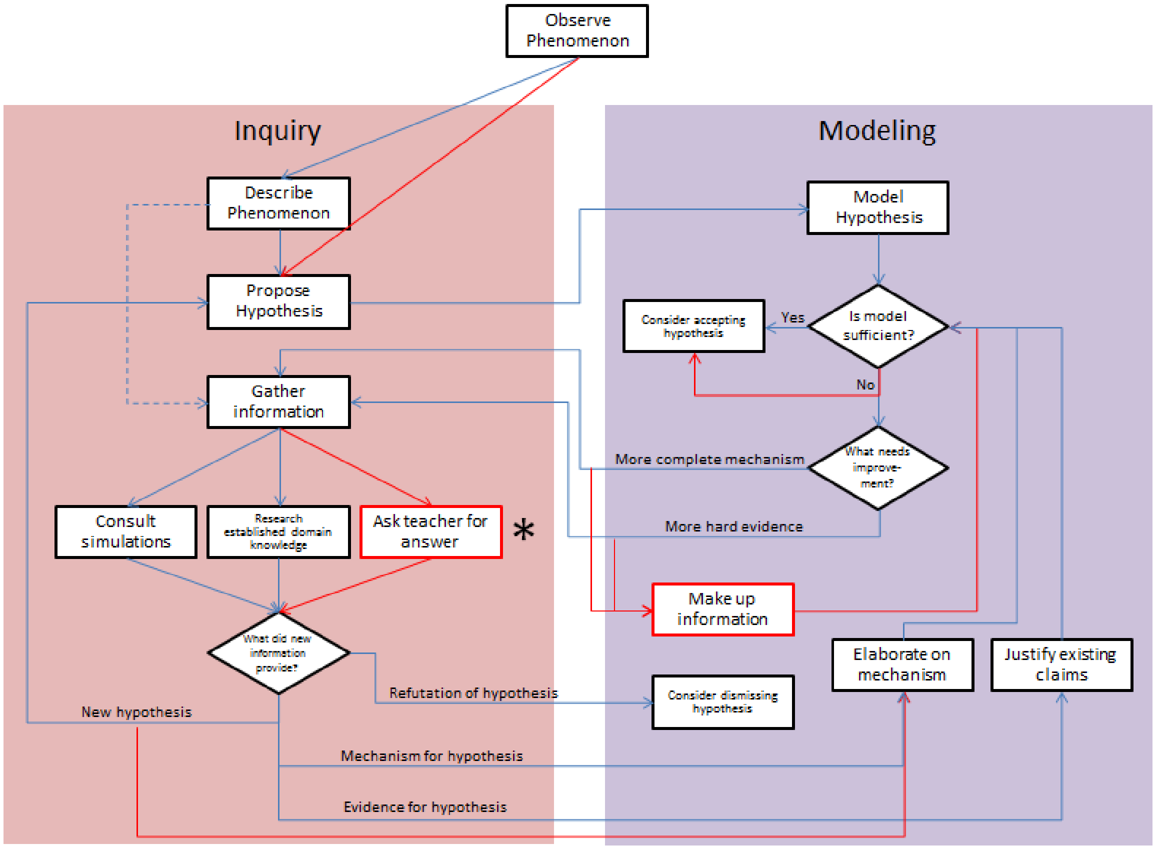
During that information gathering process, however, they also run across information on how oxygen depletion caused a similar fish kill in a different lake. They note that oxygen depletion can kill thousands of fish at once, unlike pollution. This is another new piece of information to evaluate under the 'What did new information provide?' decision, and in this case, it leads to a new hypothesis. Thus, the team now proposes a new hypothesis: oxygen depletion caused the fish kill. This leads to a new model: the rapid depletion of oxygen caused the rapid decline of the fish population. This new model is also insufficient, however, leading the team back to the information-gathering phase, now looking for evidence to support or explain how oxygen depletion led to the fish kill.

To investigate this, the team decides to ask an expert at the lake about the oxygen levels, an example of consulting data sources as a way of gathering information (in this instance, the expert is a means to access the conclusions of the underlying data). The expert tells them that after the fish kill, biologists measured the oxygen in the lake and found out that oxygen levels were dangerously low. This is a new piece of information as well, and upon evaluation, the team discerns that this piece of information provides evidence that their hypothesis is true: their hypothesis predicted low oxygen levels in the lake, and data gathered afterward corroborated that prediction. The team is now fairly confident that oxygen depletion is responsible for the fish kill, but what they do not yet understand is how exactly the oxygen depletion occurred. They bring up a simulation of a lake, and through experimentation they find that oxygen depletion is often preceded by an algal bloom: when algae populations spike, an oxygen crash often follows. This new piece of information is evaluated and determined to help provide a mechanism for their model: the drop in oxygen was caused by a rise in algal populations. Thus, they now have a three-stage model with one piece of evidence supporting it: a rise in algae populations causes a drop in oxygen concentration, which causes a drop in fish population, as evidenced in part by the observed low concentration of oxygen shortly after the fish kill.

At this stage of the process, the team of students is not done; there remain questions to be asked. For example, oxygen depletion can occur due to factors other than algal blooms; for this model to be sufficient, the team would need to supply evidence that an algal bloom was responsible for the oxygen depletion rather than stratification or the emergence of a dead zone (two alternate explanations of oxygen depletion and fish kills). Similarly, there exists a mechanism by which algal blooms lead to oxygen depletion, as well as a mechanism by which oxygen depletion leads to fish kills. A good model of this phenomenon would provide the mechanisms behind these explanation of the chain as well. Thus, the team may continue to iterate through the inquiry-driven modeling process, refining their mechanism, expanding their evidence, and improving their understanding of the system.

Errors in Inquiry-Driven Modeling

In addition to this desired process, analysis of the pilot study, earlier experience in scientific modeling education, and literature on common scientific misconceptions allowed a second model to be developed augmenting this desired process with a set of common errors students repeatedly make when attempting to participate in inquiry-driven modeling as novices. The mistakes here represent instances where students diverge from the model in order participate in generally undesirable behaviors, not instances where they depart from the model and may instead participate in other desirable inquiry processes. These are presented in Figure 5, below, based on where they occur in the modeling process.



* - Numerous other sources of information possible; only a subset is displayed here.

Figure 5: Common errors in the inquiry-driven modeling process. Red arrows denote common errors during the process.

To explore these errors, let us consider them each in the context of the investigation described previously. The examples presented here are not drawn directly from the results of the classroom intervention described later in this dissertation, but mistakes of all five kinds did occur during the intervention, inspiring the examples presented here. First, investigators may enter into the modeling process prior to coming to a strong understanding of the phenomenon they are attempting to explain; this is shown in the model as the red arrow from observing the phenomenon to proposing hypotheses. For example, a team may watch the video introducing the Lake Clara Meer fish kill and note that the video mentions visitors are staying away from the park as a result of the foul smell originating from the fish kill. This observation is not relevant to the fish kill itself, but in the absence of an articulation of the phenomenon to explain, it can be difficult to focus on the important portions of the phenomenon. This can present problems later in the modeling process as the pool of possible evidence and data is unconstrained, and models that attempt to explain multiple phenomena at once can be intractable in overall size.

A second mistake come from the resistance of students to engage in the inquiry process authentically. As described previously, students often focus more on the content knowledge and "right answer" than the process (Roll et al. 2007). As such, with a phenomenon like the fish kill in Lake Clara Meer, they demonstrate a tendency to simply ask for the answer instead of engaging in the inquiry-driven modeling process. This is marked by the red arrow from gathering information to asking the teacher for the answer. This is also a somewhat delicate error: the teacher, as an individual possessing some expert knowledge, may be a source for authentic information-gathering when students ask specific questions directed by their models. The error occurs when students instead ask general questions about the phenomenon as a whole and take the teacher's responses as sacrosanct rather than individual pieces of information to be evaluated and incorporated accordingly.

A third mistake arises when students encounter evidence that is indicative of a new hypothesis; in these instances, sometimes students will attempt to incorporate the new information into their present model without regard for its irrelevance to their model's prior claims. In the model in Figure 5, this is indicated by the red line running out of the 'New Hypothesis' exit to 'What does the new information provide?' For example, when the team in the previous example uncovered the note on oxygen depletion, they might modify their model to suggest pollution causes oxygen depletion which causes the fish to die off. However, there was no information in their investigation to suggest pollution caused oxygen depletion. Rather than begin a new explanation they instead assume the new information is part of their prior hypothesis even in the absence of data suggesting the prior hypothesis is worth pursuing.

A fourth mistake arises when students working with an insufficient explanation make up information rather than engage in information-gathering about the phenomenon; in the model in Figure 5, this is indicated by the red line running from the exits from 'What needs improvement?' to the box indicating 'Make up information'. In some cases, this may be an intentionally delinquent behavior, but other times this arises in students seemingly earnestly engaging in the process. The error is that the students mistake conjecture and hypothesis for observed fact. In the previous example, confronted with a model that proposed oxygen depletion kills off the fish, the team may suggest that the reason oxygen depletion kills off the fish is because it kills the fish's food source. When pressed, they may suggest that they believe this because it makes sense that the fish's food source dies in the absence of oxygen as well. The error here is that the team infers that if their explanation makes sense, it is therefore accurate, rather than relying on external evidence to confirm their explanation. Thus, the 'make up information' task can be seen as relying too much on conjecture rather than hard evidence.

A fifth mistake occurs when teams jump to accepting a hypothesis too quickly. At the start of the activity, teams are provided a set of model evaluation criteria that ought to be used in assessing their models. Many times, however, teams demonstrate a tendency to evaluate their model as sufficient even when it does not meet these criteria. In the above example, the team may infer that their three-phase model with the one piece of evidence is sufficient; after all, they have shown that oxygen depletion is responsible and given an idea for what may have caused it. However, there is no evidence that their proposed cause of the initial oxygen depletion is accurate, nor is there a complete mechanism indicating how the algal bloom actually led to the observed fish kill. Therefore, the model remains insufficient, and the team is mistaken in accepting it as sufficient, due either to a lack of understanding of the criteria for model acceptance or an intentional disregard for the criteria of a good explanation.

This list of mistakes is not exhaustive, nor is it systematically verified. Rather, these are some mistakes anecdotally identified during past interventions using exploratory learning environments, both in the history of this project and in other projects serving similar goals. Although the tutoring system constructed in this project aims to detect and correct these errors, it focuses more prominently on instructing students on the correct process outlined in Figure 4 on page 25.

Development of the Model of Inquiry-Driven Modeling

In developing a model of a desirable inquiry-driven modeling process, we first examined the existing literature on inquiry and modeling, paying significant attention to both the philosophy of science community and the science education community. Most significantly, we examined the work of Crawford & Stucki (1990), Schwarz et al. (2005), Nersessian (1992; 1999; 2008), Razzouk & Schute (2012), White & Frederiksen (1998), Clement (2008), Svoboda & Passmore (2013), and Darden (1998). The ideas generated by these theorists are generally compatible with an overall model of scientific modeling that involves four phases: model construction, model use, model evaluation, and model revision. In the context of inquiry, model use and evaluation take on a special form. Models are used to inform where the inquirer looks for additional information, either to corroborate and strengthen the claims of the model, to refute the claims of the model, or to elaborate on the mechanism within the model. The model is then evaluated based on how well it predicted the data that were actually founded and revised toward increasing its accuracy in the future (Schwarz & White 2005). As more new findings are accurately predicted by the model, the model's acceptability grows, and as more pieces of contradictory evidence are uncovered, its risk of being dismissed rises. This overall idea is described in several places throughout the literature, but prior to the model presented above, there did not exist a process model of inquiry-driven modeling. The closest connection to the model presented here is the existing research on instructing students in understanding scientific modeling put together by Schwarz et al. (2009), which describes a sequence for teaching modeling that builds on the construction pattern described previously, but does not operationalize modeling and inquiry into a process diagram. The model presented above is compatible with the existing accepted notions of the modeling process, including Schwarz et al. (2009) and Nersessian (2008)'s emphasis on initially "anchoring" the phenomenon, while also formalizing the process into a workflow diagram that a metacognitive tutoring system can instruct and against which students' actual behavior can be compared.

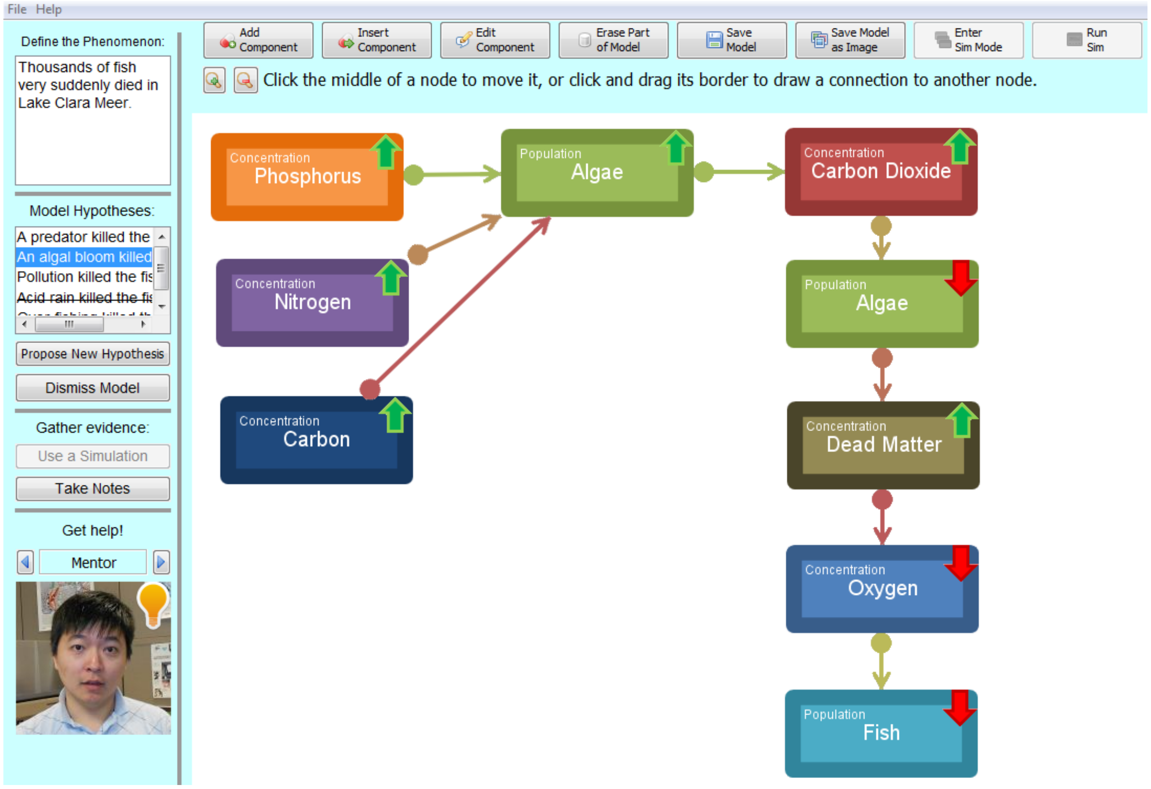
In addition to the literature on inquiry and modeling, the model of inquiry-driven modeling presented below was also developed based on several years of experience in using modeling to teach complex systems understanding in middle school classrooms. Early on, this project focused on teaching students to develop a functionally-oriented understanding of a complex ecological system (Goel et al. 2010), and early results showed success in the exploratory learning environment in increasing deep understanding of complex systems (Goel et al. 2013), providing the conceptual foundation on which MILA and MILA–T were constructed. While the project began by emphasizing the modeling element, over time it evolved to consider modeling and inquiry together (Joyner et al. 2012). During this process, a number of positive traits were identified, such as willingness to reject a previously-developed model based on new evidence. A number of mistakes in the process were identified, too, such as developing models for phenomena first and attempting to fit the evidence to them after the fact. These observations, coupled with the literature described above, formed the foundation for an initial model of inquiry-driven modeling. It is important to note that the model provided above is not intended to represent the ideal execution of the inquiry-driven modeling process, nor is it meant to present a valid representation of the way in which experts conduct scientific inquiry. Rather, it is presented as a desirable model of a way in which the process may be conducted that is attainable by the target audience of middle school students while remaining compatible with the existing literature on modeling and inquiry in scientific practice and in education.

Early forms of the tools presented in this study, MILA and MILA–T, were developed in 2012 based on preliminary modeling of the target skill from the two sources presented above. When necessary to distinguish, the early versions of these tools are referred to as MILA1 and MILA–T1, while the versions of these tools present in this dissertation are MILA2 and MILA–T2. Generally, however, this dissertation will discuss the design rationale of both versions together. MILA1 and MILA–T1 were tested in a two-week pilot study in the summer of 2012 (Joyner, Goel, & Majerich 2013). The pilot study was taught by a middle school science teacher from an Atlanta-area middle school, and the curriculum was developed by an experienced learning scientist. During this study, 16 students engaged with MILA1 with MILA–T1 active within it. Detailed log information was kept regarding students' interactions and experiences with the software and with the tutoring system. During the pilot study, students were asked to address the question: what caused the sudden massive fish kill in Lake Clara Meer? Aimed at addressing this question, students engaged in a process of inquiry and modeling to model their current understanding of the problem and use that model to dictate their continued examination of and evidence gathering for that problem. Based on analysis of these data at the conclusion of the pilot study, we adjusted the model of inquiry-driven modeling, leading to the model presented above.

Equipped with this model, this research now endeavors to create two tools. First, in order to facilitate this process, this research demands an exploratory learning environment in which students can tracked during participation in an inquiry-driven modeling task. Then, within that exploratory learning environment, a metacognitive tutoring system that guides, instructs, and apprentices students in the inquiry-driven modeling process can be developed.

MILA: Modeling & Inquiry Learning Application

As described above, MILA (**M**odeling & **I**nquiry **L**earning **A**pplication) is the latest in a line of exploratory learning environments to teach modeling and inquiry. MILA's roots lie first in the Aquarium Construction Toolkit (ACT; Goel et al. 2013), which allowed students to construct models of aquaria based on their observations of a real system in their classroom. ACT was generalized to the Ecological Modeling Toolkit (EMT; Joyner et al 2012), which supported modeling of various ecological phenomena. MILA builds on the modeling interface of EMT, but more explicitly brings in the role of inquiry in the inquiry-driven modeling process, featuring multiple tools for conducting inquiry within the environment (Joyner, Majerich, & Goel 2013). Technologically, MILA is implemented in Java to permit cross-platform deployments. Graphically, Java Swing is used to implement the majority of the user interface, while the modeling interface itself is designed from a blank canvas. In addition to Java and Java Swing, MILA also uses NetLogo simulations (Wilensky 1999) invoked from the "Use a Simulation" button on the left; students were provided with four simulations, two original simulations designed specifically for this project (Fish Breeding and Algal Blooms) and two designed by Uri Wilensky (Algae & Sunlight, Wilensky 2005; Climate Change, Tinker & Wilensky 2007). Engagement with MILA was situated in the context of a broader curriculum, and it is that curriculum which has given the software and the tutors a set of metacognitive skills to teach. It is important to note that although this learning environment represents a major portion of the ongoing research being conducted in this group, MILA is merely the backdrop for the work in this study. A screenshot of MILA is included below in Figure 6 for reference.



Thousands of fish very suddenly died in Lake Clara Meer.

A predator killed the
An algal bloom killed
Pollution killed the fis
Acid rain killed the fis

Click the middle of a node to move it, or click and drag its border to draw a connection to another node.

Figure 6: MILA, the environment in which the tutors are deployed.

CMP Modeling

The earlier exploratory learning environments developed in the history of this project supported students in constructing Structure-Behavior-Function (SBF) models of ecological phenomena. The SBF modeling language was originally created to allow artificial intelligence systems to reason about engineering and design (Goel, Rugaber, & Vattam 2009), while later efforts attempted to extend SBF modeling to educational settings (Goel et al. 2013). SBF models summarize a system in terms of three parts. The Structure is the physical arrangement of the parts in the system; the Structure of most aquatic ecosystems would simply be that the parts are mingled together within the lake or ocean. The Behavior is the processes that occur based on interactions amongst the parts of the Structure. The Function is the result of the mechanism, which can then be used as part of the Behavior for a higher-level model. For a more complete description of the results of using SBF in education, see Goel et al. 2013 and Joyner et al. 2012.

Although use of SBF modeling was shown to improve students' deep understanding of complex ecological systems (Goel et al. 2013), we also found that the vocabulary proved confusing for students and teachers alike. First, 'Behavior' most notably was regarded as the activity of a single organism, whereas in SBF behavior refers to the activity of a system as a whole. Second, 'Function' was similarly misleading; for many, the word 'function' suggested a system was designed intentionally to accomplish a certain purpose. In engineering where SBF was first used, this suggestion was accurate, but it risked major teleological misconceptions when applied to biology and ecology. In these systems, rather than designing a system to have a certain Function, we look at systems' output and build models that explain that output. Third, 'Structure' put an emphasis on the physical arrangement of parts, whereas in ecosystems the emphasis should be placed more generally on the pieces present in the system.

Thus, beginning with the later generations of EMT (Joyner et al. 2012) and extending into development of MILA, a new vocabulary was developed: Component-Mechanism-Phenomenon. Each part of this new vocabulary is directly analogous to a term in SBF. Components represent the Structure of the system, emphasizing the physical pieces of an ecological system and their properties. Mechanism represents the Behavior of the system, a sequence of developments linked together in a causal chain that results in some observable Phenomenon. Phenomenon represents what was formerly described as the Function of the system; the Phenomenon is the initial observable event that the investigator aims to explain.

To instantiate the principles of CMP and better facilitate modeling of ecological phenomena, a new modeling interface was developed for EMT and subsequently carried over to MILA. In this modeling interface, students begin with a description of the target Phenomenon to explain. This is initially written in plain text, and subsequently is articulated within the model. Students then construct a Mechanism that explains how the Phenomenon arose, grounded in the interaction of multiple Components and their properties. It is important to note that the modeling interface used in MILA instantiates the principles of CMP modeling, but this particular instantiation is not synonymous with CMP modeling. MILA implements certain features, such as management of multiple hypotheses and the inclusion of evidence within models, that are not core to the principles of CMP but rather are important for using CMP in inquiry-driven modeling of complex ecological systems.

Modeling in MILA

In order to understand how modeling in MILA provides the environment in which inquiry-driven modeling can occur, we will first explore specifically the design of MILA and its instantiation of CMP modeling. This section will first describe the overall structure of a project in MILA, then the semantics of models produced in MILA, and finally the prominent role that evidence plays in models created within MILA.

Project Structure

In MILA, all interaction takes place in the context of a project; the first task students are asked to do upon beginning engagement with the software is to create a project. The framing, organizing element behind a project is a particular Phenomenon to be explained; thus, the Phenomenon portion of CMP models is the central motivating element behind projects in MILA. After creating their project, students are asked to define the Phenomenon they are attempting to explain in a dedicated natural language input box. Each project is comprised primarily of models, each corresponding to a hypothesis for what may have caused the Phenomenon; students can also take plaintext notes during their inquiry. Each model provides a Mechanistic explanation for how a particular hypothesis may have led to the Phenomenon. Models allow students to model in parallel multiple hypotheses for what might be causing the Phenomenon; in this way, students effectively construct multiple CMP models, each with the same Phenomenon but different potential Mechanisms explaining different hypothesized causes. At any time, students can dismiss a hypothesis that they have deemed to no longer be worth pursuing; Dismissing a hypothesis crosses the hypothesis out in in the 'Model Hypotheses' box on the left, but the hypothesis can still be viewed for reference, as well as reconsidered later if additional evidence arises supporting the previously-dismissed hypothesis.

Model Semantics

Development of the specific semantics for modeling MILA was informed by prior experience with classroom interventions using ACT and EMT. One major observation during these interventions was that in the absence of constraints, students tended to put a wide variety of different types of content into their models. The resultant models lacked any consistent semantics. In designing modeling in MILA, we equipped the modeling interface with semantics that structured students' thought processes toward considering the system in terms of its Components, Mechanism, and Phenomenon.

For the purpose of explaining the semantics of models constructed in MILA, a sample model is provided below in Figure 7.

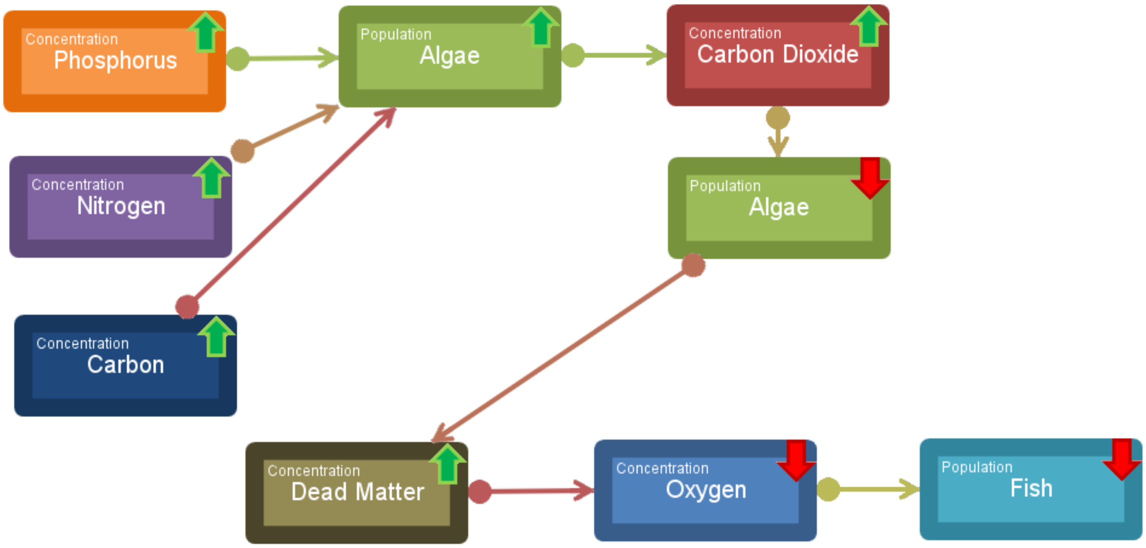


Figure 7: A sample model constructed in MILA. Note that this model was constructed by an expert researcher; examples of student models can be found in subsequent sections.

The primary objects of MILA models can be thought of as nodes and edges, although there are strong semantics at play in those nodes and edges. Each node is comprised of three parts: in the middle, the Component to which the node relates (providing the 'C' in CMP modeling), representing a physical part of the system; in the top left, a particular variable or property about that Component, representing the facet of the Component that is actually changing; and in the top right, an annotation of the direction of the change. A sample node is shown below in Figure 8; this node represents the Concentration (the variable) of Oxygen (the Component) going down.

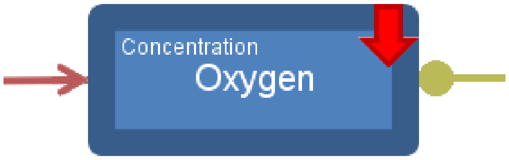


Figure 8: A node in a MILA model, representing the Concentration of Oxygen going down.

Nodes in a MILA model are not themselves Components, but rather refer to changes or alterations made to some variable about a Component. In this way, nodes represent trends, changes, or developments: the above node represents Oxygen Concentration falling, an event in a system. The Components of the models are the physical pieces of the system that are shown in the center of the nodes, but the nodes represent more than Components alone. Similarly, multiple nodes could describe the same Component; in the model in Figure 7, for example, two different nodes portray the Component 'Algae' at different points in the causal process. In other models, multiple nodes could describe different variables about the same Component; there might, for instance, exist a relationship between Fish Population and Fish Birth Rate, two variables of the same Component.

Nodes are linked together with edges, and edges represent causal relationships. In this way, the sequence of nodes, as connected by edges, represents the Mechanism, comprised of changes to Components, that students are proposing to explain the Phenomenon. An example of a short such causal chain (itself a subset of the sample model, above) is shown below in Figure 9.

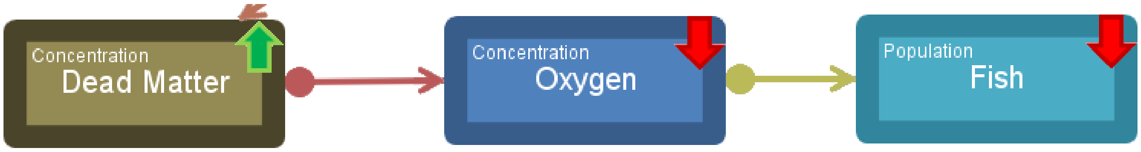


Figure 9: A portion of a model explaining how the Fish Population has dropped.

In this subset of the sample, model, the semantics of the model read as follows: the rising Concentration of Dead Matter in the system is causing the Concentration of Oxygen in the system to fall, which in turn is causing the Population of Fish in the system to drop. This is a small, linear subset of the model proposed above, but non-linear models are also possible.

Evidence in Modeling

The software, broadly, is meant in part to facilitate an authentic process of inquiry-driven modeling, and as part of that process, we aim to teach students that any propositions they make about the Mechanism of a system must be defended by some sort of Evidence; in other words, students need to explain *why* they believe the Phenomenon happened in the way they are proposing. In order to facilitate this behavior better, Evidence is actually included in the models students construct on the links they draw between nodes. Every connection between nodes is colored from red to green representing the strength of the evidence that students have currently supplied for that connection. Upon clicking on the edge, the students are provided a separate dialog into which to provide their evidence. An example of a colored edge and an example of the pop-up dialog are shown in Figure 10 and Figure 11, on pages 42 and 44, respectively.

During the previous walk-through of the inquiry-driven modeling process on page 25, we discussed how students uncover new information through the inquiry process that then influences their model. In some cases, the effect of this new information is to corroborate the claims of the model. For example, the model may make a prediction (for example, that oxygen concentration in the lake will be low). If that prediction turns out to be true, that lends credence to the model as an accurate description of the phenomenon. This process of providing evidence is crucial to the inquiry-driven modeling process. In our earlier experiments (e.g. Goel et al. 2013), one observation was that students tended to build models to explain phenomena first, and only come up with evidence for why their model was accurate later. In this way, the model's explanation was not strongly grounded in evidence. More importantly, this led to students searching for evidence solely to corroborate their model rather than grounding their model in the evidence they can provide. Thus, in MILA, evidence was elevated to a first-class object within the model itself; the mechanism that students propose is directly augmented with their evidence for why their proposed mechanism is an accurate description of the system.

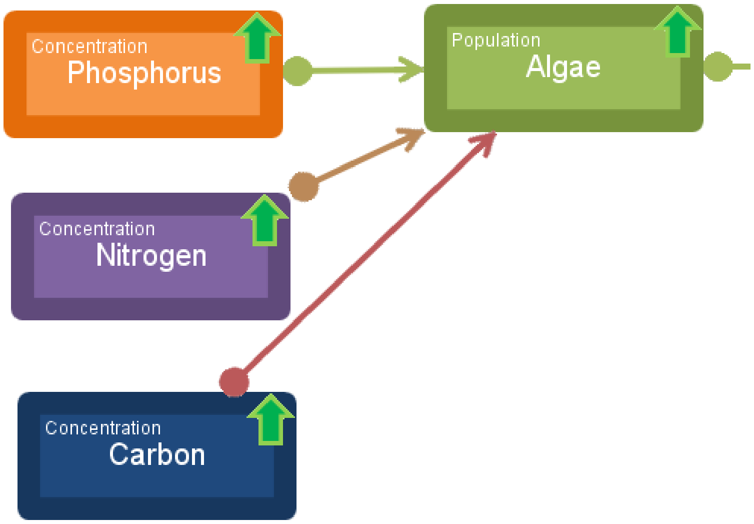


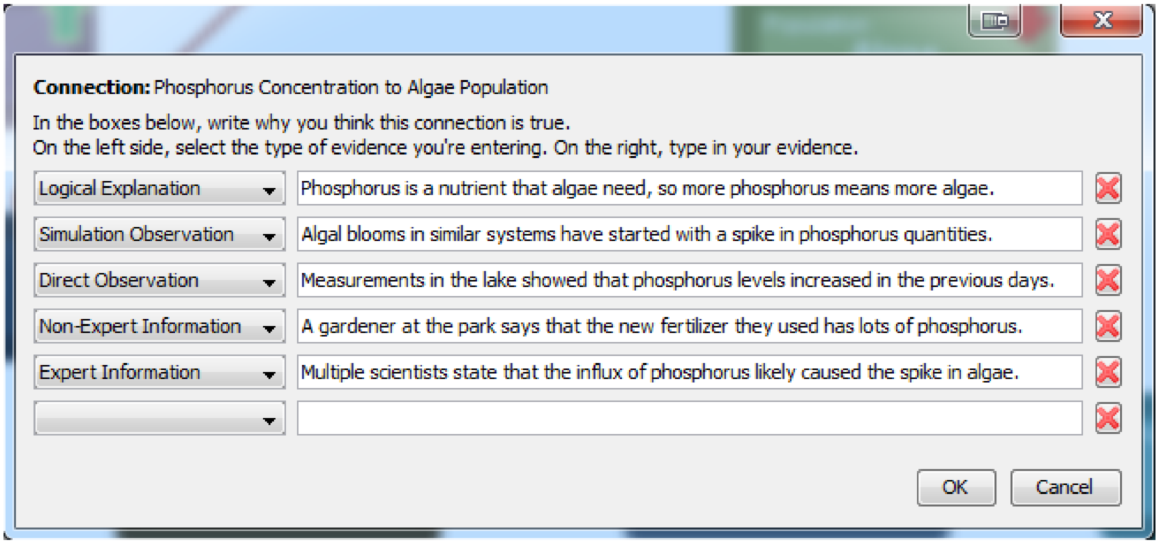
Figure 10: Several connections in a model, marked by different colors. The green connections have stronger evidence supporting them than the red connections.

When providing their evidence, students are asked to categorize their evidence using a brief categorization scheme taught elsewhere in the lesson. The categories that students can use in providing their evidence are:

- **Direct Observations**, observations on the system itself. For example, a team may observe that the fish showed no external signs of trauma, which is evidence in favor of suffocation as a cause of death.

- **Simulation Observations**, observations made while engaging with a simulation of the system. For example, the hypothetical team used previously consulted a simulation of the system and noted algae populations rose before oxygen crashes.
- **Similar System Observations**, observations made about systems in similar circumstances. For example, the hypothetical team used previously discovered oxygen depletion was responsible for a fish kill in a similar system to Lake Clara Meer.
- **Expert Information**, which includes established principles in the scientific community or testimony from a domain expert. For example, the hypothetical team used previously consulted a local expert on the lake to discover that oxygen levels were low immediately after the fish kill event.
- **Non-Expert Information**, which includes testimony from teachers, casual observers, the Internet, and other non-trustworthy sources. For example, at least one team referenced a report provided in an Internet comment wherein a person mentioned fertilizer being dumped in the park.
- **Logical Explanations**, explanations that make sense to the individual. In the example used previously, a team constructed a model that suggested oxygen depletion killed the food source for the fish because that would logically explain how oxygen depletion lowered the fish population.
- **Controlled Experiments**, experiments actually run on the system. Although less applicable in ecology given the difficult in controlling variables, students in the class conducted a controlled experiment on samples of lake water to test whether the presence of certain chemicals led to the presence of other chemicals.
- **Other Evidence**, evidence that does not necessarily fit into the previous categories.

These evidential categories are compatible with existing research on epistemic cognition (Goldin, Renken, Galyardt, & Litkowski 2014). This prior research on epistemic cognition identified five categories of evidence corresponding to the above categories of Logical Explanation, Expert Information, Non-Expert Information, Direct Observation, and Controlled Experiment. The two additional categories of evidence included here, Simulation Observation and Similar System Observation, are provided in direct correspondence to sources of data supplied during this curriculum (NetLogo simulations and articles on similar systems). These sources of evidence also build off our own research in other areas; we have examined the role of simulations in conceptual modeling (Vattam, Goel, & Rugaber 2011), as well as analogical reasoning across similar systems in design and understanding (Goel et al. 2011; Goel, Vattam, Helms, & Wiltgen 2011).



Connection: Phosphorus Concentration to Algae Population

In the boxes below, write why you think this connection is true.
On the left side, select the type of evidence you're entering. On the right, type in your evidence.

Logical Explanation	Phosphorus is a nutrient that algae need, so more phosphorus means more algae.
Simulation Observation	Algal blooms in similar systems have started with a spike in phosphorus quantities.
Direct Observation	Measurements in the lake showed that phosphorus levels increased in the previous days.
Non-Expert Information	A gardener at the park says that the new fertilizer they used has lots of phosphorus.
Expert Information	Multiple scientists state that the influx of phosphorus likely caused the spike in algae.

Figure 11: The dialog into which students can provide their evidence. On the left, students choose a category for the evidence they are providing. On the right, students write their evidence.

As students construct their models, MILA examines the evidence that students have provided for each connection, weights it to prioritize or reward more reliable evidence, and assigns a score to that connection. That score, then, is reflected in the color of the connection in the students' model. At present, a weighting or prioritization of evidence has not been identified by either the epistemic cognition community or researchers on the philosophy of science. For the purposes of this project, a simple weighting scheme was developed as a starting point to instruct students on the value of stronger types of evidence. Expert Information and Controlled Experiments were weighted the highest at three points; Simulation Observations and Similar System Observations were weighted at two points; and Logical Explanations, Non-Expert Information, and Direct Observations were weighted at one point each. This weighting scheme is not proposed as an optimal prioritization of evidence, but rather as one possible faithful recreation of the strength of different kinds of evidence uncovered by the epistemic cognition community. This weighting scheme was used to provide students feedback on the strength of their models as well as to evaluate model strength in the subsequent analysis in Chapter 4; however, the accuracy of this scheme relies on students correctly coding the category for the pieces of evidence they supply. Because there is no way to ensure this accuracy during interaction with the software, results were coded during the analysis phase to examine whether students were correctly categorizing the evidence they provided and adjust the scoring for different models accordingly.

MILA–T: The Metacognitive Tutoring System

As described previously, two versions of MILA–T have existed: MILA-T1, used in a pilot study in 2012, and MILA-T2, used in the primary experiment for this dissertation. In MILA–T1, four tutors were used: a Guide, a Critic, a Mentor, and an Interviewer. These tutors used some relatively simple production rules to generate feedback for the students based on the current day of the summer camp, the classroom context, and, to a limited extent, the content of their model. Based on observations from the pilot study and the creation of the aforementioned model of inquiry-driven modeling, the four tutors were revised to give much more sophisticated advice, emphasize metacognitive skills, and maintain an underlying model of student efficacy. Toward this end, a fifth tutor was also created: an Observer, to keep track of students' progress and update an underlying profile of teams' ability for use by the other tutors. This batch of tutors, MILA–T2, was used in the project described in this dissertation. Unless otherwise noted, 'MILA–T' in this dissertation refers to MILA–T2.

Broadly, all five tutors are characterized in some way as mimicking a functional role that a teacher plays in a classroom, loosely corresponding to the functional roles proposed in the Grasha model of teaching styles (Grasha 1996) and leveraging Shute's notion of different qualitative types of formative feedback (Shute 2008). This serves several purposes. First, these tutors are intended to provide the kinds of feedback that would be provided by a teacher or human tutor, and therefore it is useful to consider the functional roles that these individuals play in choosing the roles for these tutors (Merrill, Reiser, Ranney, & Trafton 1992). Second, examination of the functional roles played by teachers and human tutors gives insight into the types of and prompts for feedback that tutors may provide (VanLehn et al. 2003; Olney et al. 2012; D'Mello & Graesser 2013). Third, in our deployments, these tutors are used in a classroom environment, allowing the tutors to offload some of the demand on teachers onto the software and supporting the teachers' efforts to provide higher-level feedback. The analysis in Chapter 5 suggests, in fact, that this process of offloading functional roles from the teachers may be the primary mechanism behind the learning gains seen in the Experimental group of students. In addition to structuring the tutors around these functional roles, each tutor is also supplied with a human name and a human face to attempt to leverage the benefits seen in prior research toward spawning empathy between the student and the tutoring system (Lepper & Chabay 1988; Johnson, Rickel & Lester 2000; Azevedo et al. 2009; Lee & Ko 2011).

On Metacognition

As described at the beginning of this dissertation, the goal of this tutoring system is to teach students the metacognitive process of inquiry-driven modeling. However, the functional roles played by the tutors do not inherently connect to metacognition; it would be possible to design a similar tutoring system that teaches cognitive skills or content knowledge. Rather, the way in which MILA–T teaches metacognition is in the content of the tutors. The tutors teach metacognition in three primary ways: by shifting students' focus to metacognitive skills, by instructing proper metacognitive abilities, and by demonstrating metacognitive thought themselves.

As mentioned previously, one of the key challenges with teaching metacognition is instructing students to value the metacognitive process itself rather than the content knowledge. Our past interventions showed students had a strong proclivity to value the correct answer above all else, going so far as to actively eschew the desired process if they thought it would give them the right answer. This dynamic informed the selection of the phenomenon used in this intervention: the correct explanation of the fish kill in Lake Clara Meer is not conclusively known. Thus, the first way in which the tutoring system emphasizes metacognition is by attempting to shift students' focus to metacognitive processes rather than straightforward content knowledge. For example, the tutors anticipate certain undesirable questions students might ask, like "What really did cause the fish kill?", and reply by emphasizing the process by which they ought to answer that question. The tutors also ask students questions that explicitly encourage students to reflect on why they made the decisions that they made rather than just the content of the decision.

The second way in which the tutors of MILA–T target metacognition is with the actual content that they provide to students. The feedback that tutors have available is all in service of metacognition and inquiry-driven modeling; the tutors frequently suggest to students how to and when to engage in metacognitive processes like planning their next move, evaluating and monitoring their current understanding, incorporating new knowledge into their prior understanding, and evaluating their prior understanding against that new knowledge. Thus, not only do the tutors encourage students to emphasize their own metacognition, but they also instruct students on how exactly they should be thinking about the inquiry and modeling task. In this way, the tutors do not only teach students *to* pay attention to how they are thinking about the problem, but also actually *how* they should be thinking about the problem.

In service of this attempt to directly and explicitly instruct the desired metacognitive process, the tutoring system also engages in teaching by example. In addition to giving explicit suggestions on what the team should do next, the tutors also describe the proper decision-making process for different circumstances. For example, an error described in the model shown in Figure 5 on page 29 is that students are not always sure of how to recover from and respond to evidence that contradicts their current hypotheses. In anticipation of this error, the tutors are equipped with responses that give examples of how to respond to those situations. In providing these answers, however, the tutors are most concerned with instructing the thought process underlying those responses, rather than the output of the response itself. In this way, the tutors emphasize metacognition by emphasizing it, instructing it, and demonstrating it.

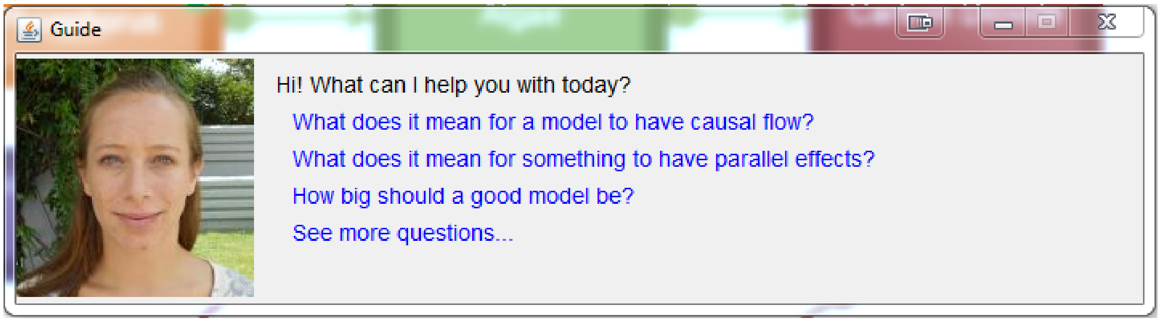
In addition to targeting students' metacognition, the tutors of MILA–T can also be described as metacognitive in that they themselves are thinking about thinking. Metacognition, as described previously, is defined as thought about thought, and the tutors of MILA–T are explicitly trying to teach thought. Thus, the tutors can be described as metacognitive in that they reason over students' own cognitive and metacognitive thought. This is significant in two ways. First, it highlights the explicit limitations placed on the tutoring system. They do not, for example, check the biological accuracy of the models that students construct because biological accuracy is not the skill they aim to teach; instead, they aim to teach the thought process that will ultimately lead to biological accuracy, not only with known systems but also with as-yet unexplained phenomena. Secondly, it emphasizes the apprenticeship nature woven into many of the tutors' responses. The tutors are frequently informing students what decisions they would make if they were in the same situation and what thought processes would go into those decisions. In this way, the metacognitive tutors attempt to make their own metacognition explicit to the students in order to both emphasize and demonstrate a desirable metacognitive process.

The Five Tutors

MILA–T is comprised of five tutors: Gabriel the Guide, Craig the Critic, Mercer the Mentor, Isla the Interviewer, and the nameless Observer. For each of these tutors, this section will describe the functional role of a teacher that the tutor mimics, the detailed functional information by which the tutor mimics that role, a set of production rules emblematic of the tutor's overall operation, and the way in which the tutor attempts to address students' metacognitive processes of inquiry-driven modeling.

The Guide

The Guide, named Gabriel in the software, offers pre-written questions to which the team might want the answer based on their current models, past modeling behaviors, and the current classroom activity. Figure 12 below shows the Guide offering a selection of questions to the team.



Hi! What can I help you with today?

What does it mean for a model to have causal flow?

What does it mean for something to have parallel effects?

How big should a good model be?

See more questions...

Figure 12: The Guide, offering a selection of questions chosen based on its production rules.

<u>Functional Role</u>

The Guide simulates the functional role played by the teacher of answering questions that the students bring up. In this situation, the student initiates asking a question, unprompted by the teacher. As such, the Guide does not interrupt the team to offer questions, but rather offers questions only when the team initiates interaction. Interaction with the Guide can resemble a conversation; the questions that teams ask can lead to additional questions and multiple layers of textual feedback, mimicking an ongoing conversation between a teacher or human tutor and the student. For example, while using simulations, the team might question the relevance of the activity to the modeling they have been doing in the past; they would then talk to the Guide, who would offer the question, "How do simulations help us build scientific models?" Upon choosing this question, the Guide would offer guidance on how to manipulate a simulation in a way that informs the team's model, or how their model can guide their usage of the simulation. The Guide can be thought of as mimicking the Expert and Delegator styles from Grasha's model of teaching styles (Grasha 1996).

Detailed Functional Information

Figure 13 below shows the input that the Guide receives and the output that she returns in response. The Guide is first prompted by the team's request for help. She then checks the current state of the Observer's profile of the team's ability, the current state of the team's models of the phenomenon, and the current classroom context to develop a list of questions to offer to the team. Each question may lead to an additional set of questions (for example, a question on software interaction may lead to individual questions about adding components, adding connections, deleting components, etc.) or an answer, presented in text.

The Guide only acts when initiated by the team. Teams consult the Guide by clicking the avatar in the bottom left corner of the MILA window, as shown in Figure 6 on page 36. When teams click the Guide, the Guide runs through a series of production rules in order to choose which questions to offer. Factors that are considered in the production rules include: the current structure of the team's model; the current classroom context, including what day of the curriculum teams are currently on; the team's efficacy with modeling and inquiry according to the Observer (as described under the section "The Observer" on page 67); and the recent tasks in which the team has engaged. Iterating through these production rules, the Guide compiles a list of questions that may be appropriate to the team at the given time and prioritizes them. Then, the Guide appears to offer the prioritized list of questions from which the team can select a question to ask.

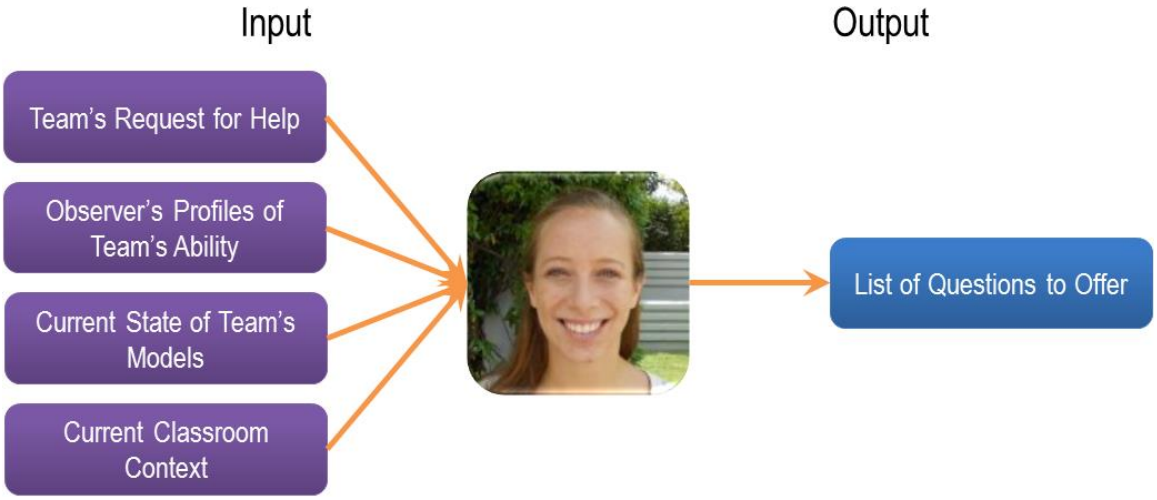

Input
Output
Team's Request for Help
Observer's Profiles of Team's Ability
Current State of Team's Models
Current Classroom Context
List of Questions to Offer

Figure 13: The inputs and outputs for Gabriel the Guide. The Guide takes in the team's request for help, checks the information available to her, and provides a list of questions suited to the team's current ability and context.

Detailed Production Rules

Below are three of the Guide's production rules of varying complexity. Note that the questions listed at the left have corresponding answers that are not shown here. Full rules can be found at dilab.gatech.edu/mila/

Table 3: Early in the project, the Guide supplies help at the software interaction level to prevent the teacher from having to explain the basic function of the software multiple times.

Guide Rule #1	
Conditions	Actions
IF: This is the first usage of the software; OR: The team has not yet created anything in the software.	THEN: Add to question list, "How do I create MILA Models?".

Table 4: After the team has demonstrated some minimum competency with the software interaction elements, the Guide shifts to prioritizing questions that instruct the team on why such a greater emphasis is being placed on evidence by connecting them to the broad

Guide Rule #2	
Conditions	**Actions**
IF: The team has begun to add evidential justifications to their models; OR: Evidence has been introduced; AND: The team has not yet reached efficacy with evidential justifications according to the Observer's profile. OR: The team's current model is very large but lacks any evidential justifications.	THEN: Add to question list, "What does evidence mean?"; AND: Add to question list, "What are the different types of evidence?"; AND: Add to question list, "How should evidence be used in a model?"; AND: Add to question list, "What is the importance of evidence in science?".

Table 5: Once the team has engaged in a significant amount of modeling and developed multiple decent hypotheses and explanations, the Guide infers that the team will need to start pruning away some explanations soon.

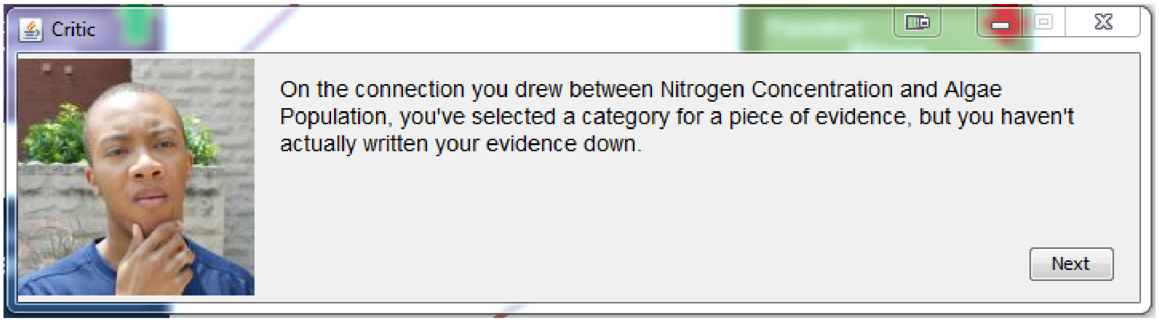
Guide Rule #3	
Conditions	**Actions**
IF: The team has multiple models meeting a minimum criteria of development; OR: The team has demonstrated an intermediate efficacy with inquiry; AND: The team has demonstrated a basic efficacy with modeling; OR: The team is on the last day of the lesson; AND: The team has multiple models under consideration.	THEN: Add to question list, "When should I consider dismissing a hypothesis?" question. AND: Add to question list, "How do I know which hypothesis to dismiss?" AND: Add to question list, "Does dismissing a hypothesis mean I was wrong?"

On Metacognition

The Guide supports teams' metacognitive inquiry-driven modeling process in several ways. First, the structure of the questions themselves is intended to prompt the team to attend to metacognitive processes in the first place; while the Guide attempts to anticipate the questions teams will ask, she also anticipates the questions teams *should* ask if they are engaging in the desirable metacognitive process. Secondly, the Guide's feedback nearly always evokes the team's metacognitive processes in its prescriptions; even on questions that are concerned primarily with the technical usage of the software (such as how to add nodes), the Guide touches on the way in which teams should think about the models they are producing and the role they play as external representations of their understanding. Third, the Guide is consistently concerned with students' dispositional approach to the process; the Guide reminds teams in several places that perceived failures like dismissing a failed model are actually positive steps toward an understanding of the phenomenon.

The Critic

The Critic, like the Guide, responds to teams' requests for feedback, only acting when the team clicks on the Critic's avatar in the bottom left. When consulted, the Critic gives one piece of feedback on the team's current model. The Critic is shown in Figure 14 below.



On the connection you drew between Nitrogen Concentration and Algae Population, you've selected a category for a piece of evidence, but you haven't actually written your evidence down.

Figure 14: The Critic, providing a piece of feedback based on the current model.

Functional Role

The Critic simulates the functional role of responding to students' requests for feedback on their work, and as such, the Critic is almost exclusively concerned with the actual contents of the team's models. In a classroom, during paper exercises and other activities, the teacher may be available for students to approach and ask for feedback on their work at present. The teacher then gives feedback on the work shown. As such, similar to the Guide, the Critic waits for teams to request feedback, and gives feedback primarily on the current status of what they have produced (although the Critic does allude to teams' interaction histories in some instances). For example, when a team believes they have constructed a satisfactory model (as represented in the model of inquiry-driven modeling presented earlier), they may ask the Critic if their model is satisfactory. The Critic would typically then offer an area where the model can be further improved. The Critic can be thought of as mimicking the Formal Authority style from Grasha's model of teaching styles (Grasha 1996).

Detailed Functional Information

Figure 15 below shows the input that the Critic receives and the output that he returns in response. The Critic is first prompted by the team's request for help. He then checks the current state of the Observer's profile of the team's ability (as described under the section "The Observer" on page 67), the current state of the team's models of the phenomenon, and the current classroom context to develop a list of possible pieces of feedback to provide to teams; in this way, the percepts that the Critic checks are the same as those that the Guide checks, but the decisions made on the basis of those percepts differ. After developing a list of possible pieces of feedback, the Critic randomly chooses one to provide with an accompanying change in facial expression.

Like the Guide, the Critic prepares feedback only at the time when the team initiates communication. At that time, the Critic runs through a sequence of production rules reflecting various desirable criteria and common errors that might be found in a team's model. These production rules are prioritized according to several criteria, including complexity, necessity to future critiques, and past modeling performance. For example, the Critic checks on the strength of evidence the team has provided for their model, but it is not helpful to provide feedback on evidence strength when the team has not even drawn a complete model. Similarly, if there is a glaring problem with a model, such as the presence of an orphaned node or evidence with no textual annotation, it is not prudent to give advanced critiques rather than fixing simpler problems.

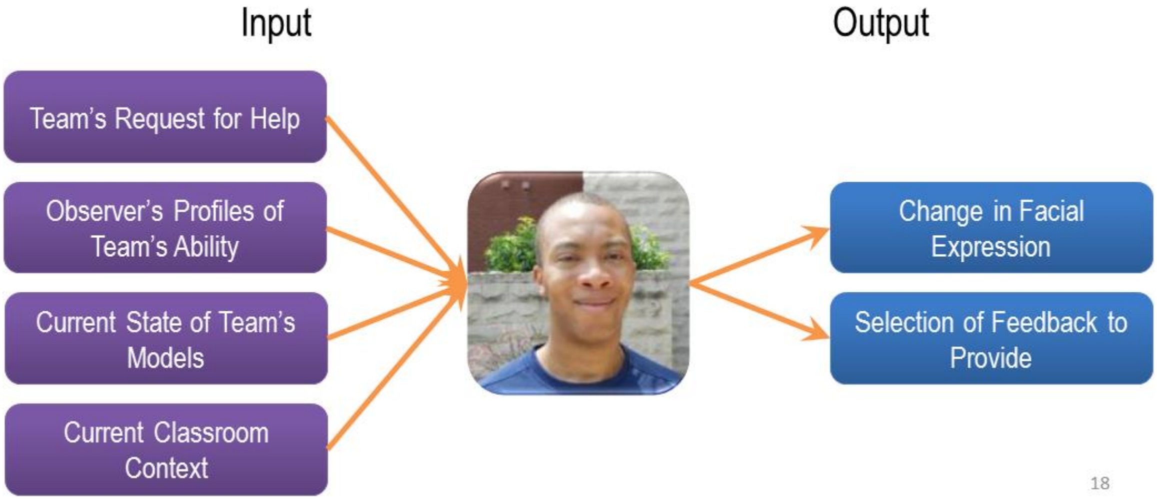

Input
Output
Team's Request for Help
Observer's Profiles of Team's Ability
Current State of Team's Models
Current Classroom Context
Change in Facial Expression
Selection of Feedback to Provide
18

Figure 15: The inputs and outputs for Craig the Critic. The Critic takes in the team's request for help, checks the information available to him, develops a list of possible pieces of feedback to provide, and randomly selects one.

As the Critic checks its production rules, he gathers relevant feedback to provide to the team, similar to the way in which the Guide gathers questions. Once the Critic reaches specific thresholds, however, he ceases checking for additional feedback to provide and randomly selects one of the pieces of feedback he has gathered to provide to the team. These thresholds are established primarily by complexity: the Critic will stop gathering additional possible pieces of feedback once he has either gathered a useful amount of potential feedback to offer in the same relative range of complexity, or if it establishes that there is a large gap in the complexity of feedback he could offer. In this way, the Critic can typically offer the team multiple pieces of feedback at the same time if they choose to continue requesting it, preventing the Critic from getting hung up on a piece of feedback that the team cannot incorporate yet or that was erroneously selected.

Detailed Production Rules

Below are three examples of specific production rules used by the Critic of varying levels of complexity. Full rules can be found at dilab.gatech.edu/mila/

Table 6: Early on, the Critic detects one of the fundamental errors from the original inquiry-driven modeling process: teams beginning to model their hypothesis before articulating the phenomenon that they are explaining.

Critic Rule #1	
Conditions	**Actions**
IF: The team has not yet written a phenomenon definition; AND: The team has already begun creating models.	THEN: Add to the advice selection pool, "You've created some models, but you have not yet written a description of your phenomenon. Remember, it is very important to have a strong idea of what you are trying to explain before you start explaining it!"

Table 7: Later in the modeling process, the Critic will look at the team's evidence. If the model's evidence is weak, the Critic checks the model of the team's ability to see if they have demonstrated understanding of evidence earlier in the project. If not, the Critic chooses to give them feedback on the weakness of their evidence.

Critic Rule #2	
Conditions	**Actions**
IF: The team's model currently uses only weak evidence; AND: The team has not demonstrated sufficient intermediate competence with modeling, according to the Observer. OR: The team has not demonstrated sufficient intermediate competence with inquiry, according to the Observer.	THEN: Add to the advice selection pool, "So far, you are focusing on some of the weaker kinds of evidence, like Non-Expert Information and Logical Explanations. These can be good places to get started, but try to look for stronger evidence to support your ideas. A convincing model needs very strong evidence to support it."

Table 8: Once the team has a model that meets the earlier, simpler criteria, the Critic starts to set higher standards for how strong the evidence behind a model ought to be.

Critic Rule #3	
Conditions	**Actions**
IF: There exists an edge in the team's model that has a strength score under 5.	THEN: Add to the advice selection pool, "Your connection between [Start Node] and [End Node] has decent evidence on it, but it is not yet completely convincing. Try to solidify it with some strong scientific theories or direct confirming observations."

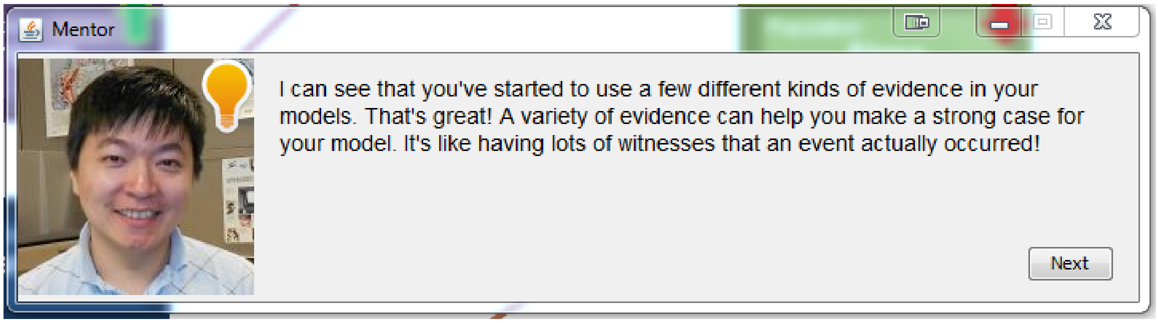
Note that the production rules for the Critic appear simple because implicit in the production rule is the knowledge that all previous production rules have not been added to the pool. For example, this final production rule is one of the final pieces of advice that teams can receive from the Critic on their models, representing the most advanced finishing touches that the Critic can suggest. Implicit in the selection of this rule is that basic errors and simpler feedback were not needed at this stage; if they had been, the Critic would have ceased looking for new feedback to provide before reaching this production rule.

On Metacognition

The primary role the Critic plays with regard to students' metacognitive abilities is the demonstration of the proper way to reason over a model. While the Critic typically gives its initial feedback in terms of a direct correction or improvement that can be made on the model, such as including Variables on Components or justifying connections with significant evidence, he always moves on to contextualize that within the thought process that that action supports. For example, providing evidence is a way of reasoning over the level of confidence that the team has in the model, so when the Critic is advising the team to add more evidence, he connects that superficial suggestion with the deeper notion of confidence more indicative of self-reflection on the model's strength. In this way, the Critic connects an observable behavior and repeatable skill with a metacognitive process.

The Mentor

Unlike the Guide and the Critic, the Mentor does not wait for the team to consult him; instead, the Mentor can interrupt the team at any time to provide feedback. When the Mentor's avatar changes to one with a light bulb, as currently shown in Figure 16, the team knows the Mentor has feedback for them. The light bulb remains until the team clicks and receives the feedback or until the prompt for the feedback is resolved without the Mentor's intervention.



I can see that you've started to use a few different kinds of evidence in your models. That's great! A variety of evidence can help you make a strong case for your model. It's like having lots of witnesses that an event actually occurred!

Figure 16: The Mentor, making a positive note on the current team's variety of evidence.

Functional Role

In a classroom or a tutoring session, the teacher or human tutor typically observes the students while they engage in the activity. For teachers, they may move around the room, looking over the shoulder of students; for human tutors, they observe while students solve problems on their own. When teachers or human tutors see the student do something that demands feedback, they provide it immediately. This feedback may come with various purposes: it may be corrective when the teacher observes a mistake; it may be supportive when the teacher observes the students solving something correctly or engaging in proper reasoning; it may be expounding when the teacher witnesses a "teachable moment" where they may generalize students' actions to a broader principle; it may be connective when the teacher can link a students' activity to a related concept; or it may be of one of several other distinct qualitative types. The important criteria for any of this feedback are that it is provided immediately and spontaneously in the context of the activity. For example, immediately after demonstrating an understanding of the role of invisible substances in their model, the Mentor might give feedback connecting the importance of such invisible substances to other processes as well; alternatively, if the team has constructed complex models without any attention to invisible substances, the Mentor might interrupt and remind teams of the role these things can play. The Mentor can be thought of as mimicking the Personal Model style from Grasha's model of teaching styles (Grasha 1996).

Detailed Functional Information

Figure 17 below shows the input that the Mentor receives and the output that he returns in response. The Mentor checks his percepts on a timed interval. Initially, he checks the current status of the Observer's profile of the team's performance (as described under the section "The Observer" on page 67), followed by the team's current models and the current context in the classroom. Using this, he builds a list of possible pieces of feedback to give to the team. With this list, he then checks the team's history of interaction with the Mentor and adjusts or dismisses feedback he has already provided in the past. If any feedback remains and if he has not interrupted the team recently, he decides to interact, illuminates his light bulb, and changes his facial expression.

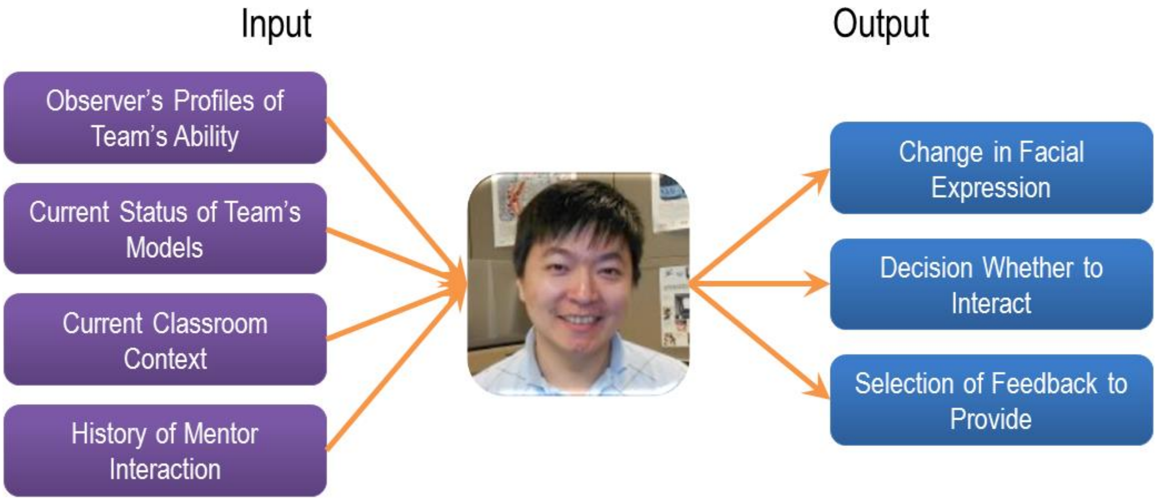

Input
Output
Observer's Profiles of Team's Ability
Current Status of Team's Models
Current Classroom Context
History of Mentor Interaction
Change in Facial Expression
Decision Whether to Interact
Selection of Feedback to Provide

Figure 17: The inputs and outputs for Mercer the Mentor. The Mentor regularly checks the percepts available to him, selects feedback to provide, and alerts the team that feedback is available.

The Mentor checks his production rules on a timed interval; the interval is changed depending on the team's activities and interaction with the Mentor and other tutors. Every time the Mentor checks his rules, he proceeds through a prioritized list and stops checking once he finds a single production rule for which the conditions are met. At that point, the feedback associated with the rule is processed and the Mentor's avatar is updated to reflect the presence of feedback: the selection panel automatically switches to the Mentor and a light bulb appears on the avatar. After this has occurred, the Mentor will continue to check his production rules on the same interval; if the conditions that prompted that feedback to be available cease to be met (for example, the team fixed the error that the Mentor was going to correct on their own), the Mentor will respond accordingly and rescind that feedback. Similarly, if higher-priority feedback becomes available, the Mentor will switch to the higher-priority feedback.

When the team clicks the Mentor while feedback is available, the Mentor's box will appear and deliver that feedback. The Mentor's feedback is all text-based, and at different points reflects the different qualitative types of feedback discussed previously. Upon viewing the feedback, the Mentor marks that feedback as 'read' permanently for the given group; even though in many cases the conditions for that feedback will continue to be met (especially when the Mentor is supporting or confirming a decision the team has made), that feedback will not appear again. For certain types of feedback, especially corrective feedback, the Mentor may comment on the same conditions in multiple ways if the conditions continue to be met, but he will not provide the exact same feedback more than once. If feedback is triggered but not read (because the conditions ceased to be true, higher-priority feedback became available, or the team closed the software before consulting the Mentor), the feedback can still be triggered and provided. Finally, if team clicks the Mentor when no feedback is available, the Mentor delivers a simple message that no feedback is currently available.

The timed interval under which the Mentor checks his production rules is variable; every time teams receive feedback from the Mentor, the time until the Mentor checks his production rules again is modified to prevent feedback from being provided too frequently. When corrective feedback is provided and additional corrective feedback is available for those conditions, the interval is modified to ensure that the team has sufficient time to correct the error before being prompted again. After the interval has passed, the Mentor returns to his regular timing (typically ten seconds). Lastly, the Mentor is also in certain places directly alerted by MILA to check his production rules after certain actions and under certain conditions; for example, the first day of the lesson, the Mentor immediately checks his production rules after several team actions to give precise feedback in the form of a tutorial. The immediate reactions by the Mentor are intended to make clear to the team what action of theirs prompted the Mentor to respond.

Detailed Production Rules

Below are three examples of specific production rules used by the Mentor of varying levels of complexity. Full rules can be found at dilab.gatech.edu/mila/

Table 9: The first of several rules that guide teams through the basics of using the software. Teams are immediately provided with feedback to create a model, add a node, add a second node, draw a connection, and insert a node into the middle of that connection. Each action triggers the Mentor to provide feedback for the teams pointing toward the next action.

Mentor Rule #1	
Conditions	**Actions**
IF: The team has just opened the project for the first time.	THEN: Make available feedback saying, "Welcome to MILA! To get started, go ahead and write a precise description of what you are trying to explain in the Phenomenon box. Remember, an important part of the scientific process is to start with a very strong idea of the phenomenon you are attempting to explain.

Table 10: A rule that reflects the Mentor's reasoning in accepting the appropriateness of the feedback, ensuring that the feedback is directly temporally relevant (the first criteria), appropriate for the team's current level of ability (the second criteria), and non-repetitive (the third criteria).

Mentor Rule #2	
Conditions	**Actions**
IF: The team has recently dismissed one of their models; AND: The team had not yet demonstrated proficiency with proposing and dismissing models according to the Observer's profile of the team; AND: The team has not yet received positive feedback on dismissing models.	THEN: Make feedback available, "I see you've dismissed one of your initial hypotheses. Well done! Proposing and then ruling out hypotheses is an important part of science. It's crucial to reflect on your ideas and understand when you have disproven an earlier hypothesis."

Table 11: An alternate rule provided when the team has not yet demonstrated mastery of a concept. The paired feedback formed by Mentor Rules #2 and #3 demonstrates a common structure in the Mentor's feedback. In many places, the Mentor gives positive feedback when teams demonstrate a certain skill or action; if the teams do not demonstrate that skill before certain other requirements are met, the Mentor provides guiding feedback to consider that skill.

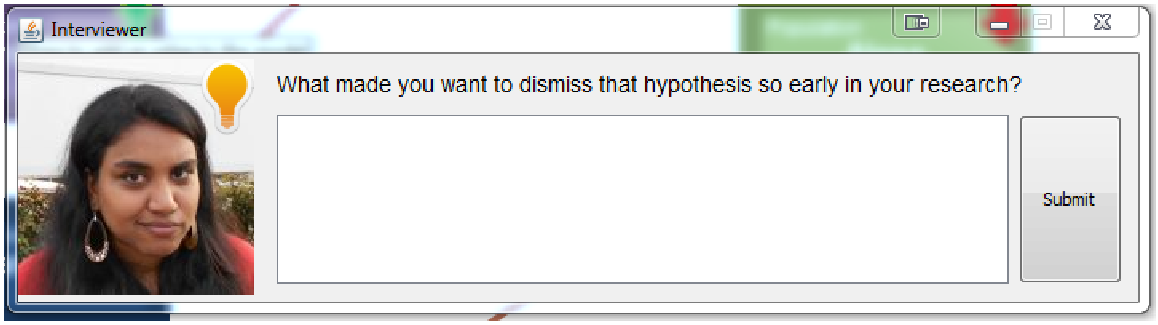
Mentor Rule #3	
Conditions	**Actions**
IF: The team has not yet dismissed a model; AND: The team has multiple models currently available in their project; AND: The team is currently on one of the later days of the lesson. OR: One model is significantly more-developed than the others.	THEN: Make feedback available, "It looks like you haven't dismissed any of your models yet. It might be time to consider doing so. Consider taking a look at all your models and narrowing down which are the strongest, or if you can't tell, investigate them all equally until a better one emerges. Remember, dismissing prior hypotheses is an important part of the scientific process. If you need more information on when and what to dismiss, check with the Guide."

On Metacognition

With regard to students' metacognitive skills, the Mentor plays at least two primary roles. First of all, similar to the Guide, the Mentor aims to attune students to the importance of metacognitive skills through repeated mention and explicit acknowledgement. The majority of the behaviors that the Mentor replies to are proxies for an underlying metacognitive ability; for example, while dismissing a model is a superficial behavior within the software consisting simply of clicking a button, the Mentor uses this behavior as a proxy for the willingness to dismiss failed hypotheses. Although this might not be an accurate proxy in all instances, the Mentor still seizes the opportunity to contextualize the behavior in a metacognitive skill; thus, even if the identification was mistaken, the Mentor still connects a recent action to the desirable metacognitive skill. Secondly, in line with the second and third Mentor rules above, the Mentor is also responsible for attempting to bring students into line with the model of inquiry-driven modeling outlined in Figure 4 on page 25; he does this by observing and encouraging the desirable behaviors seen in that figure, while attempting also to catch and correct the errors noted in Figure 5 on page 29. He also detects more abstractly lack of performance of the behaviors in Figure 4, even when absence of the behavior does not cross over into the errors in Figure 5. Toward this end, the Mentor leverages his position as part of a broader curriculum on inquiry-driven modeling; as part of perceiving the current classroom context, the Mentor perceives the time by which teams are expected to demonstrate certain skills. If the team has not yet demonstrated a skill by the threshold identified by the curriculum, the Mentor alerts the team to the presence of that skill, describes why he has not yet identified it in the team's process, and explains the value of the skill. Thus, the Mentor relies significantly on the curriculum to provide a trajectory against which to track the team's inquiry-driven modeling ability.

The Interviewer

The Interviewer, shown in Figure 18 below, is responsible for asking teams questions that they can answer in plain text. The answers are then stored in the team's project for later perusal by a teacher or researcher. The Interviewer generally asks teams questions that the scientific process dictates they ought to learn to ask themselves.



What made you want to dismiss that hypothesis so early in your research?

Figure 18: The Interviewer, asking the teams to reflect on their early dismissal of a hypothesis.

Functional Role

One important role for teachers and human tutors, especially with regard to developing metacognitive skills, is to probe and prompt the team to consider their own thought process in their decisions and behaviors. At a basic level this serves to help the teacher or tutor understand the team's current level of ability, but at a deeper level, teachers and tutors are often asking students questions that the student should consider themselves during the activity. For example, in Algebra, a teacher might ask a student why they divided both sides by 'x'. The student would then reflect on their thought process and, sometimes, realize that their thought process was faulty. In modeling, a teacher or tutor might ask a student why they dismissed a certain model; this question allows the teacher or tutor to probe the student's reasoning and correct it if necessary, but also demonstrates to students that they ought to have an explanation for that decision during their own reasoning as well. The Interviewer, for example, asks exactly this question: when teams dismiss models, the Interviewer asks them to explain their reasoning, calling special attention to certain instances of model dismissal (such as dismissing an early, unexplored hypothesis or dismissing a thorough, well-defended model). The Interviewer can also be thought of as mimicking the Facilitator style from Grasha's model of teaching styles (Grasha 1996).

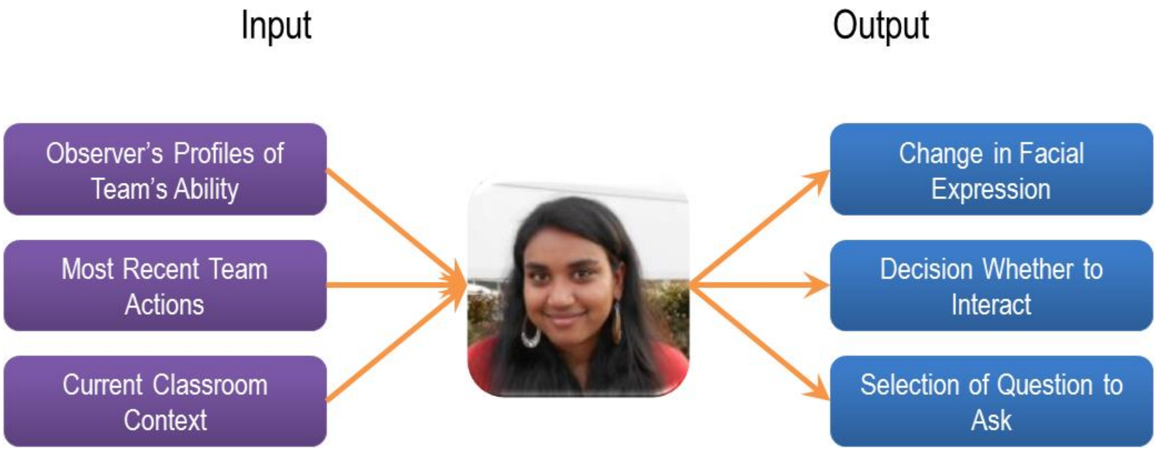

Input
Output
Observer's Profiles of Team's Ability
Most Recent Team Actions
Current Classroom Context
Change in Facial Expression
Decision Whether to Interact
Selection of Question to Ask

Figure 19: The inputs and outputs for Isla the Interviewer. The Interviewer asks primarily about specific actions that the team takes in the software, and thus checks her percepts after every action and responds when she finds an action for which she has a question.

Detailed Functional Information

Figure 19 above shows the input that the Interviewer receives and the output that she returns in response. The Interviewer checks her percepts every time the team takes an action because her questions are typically targeted at reflecting on specific actions within the inquiry-driven modeling process. She checks their most recent interaction with the Observer's current profile of the team's ability in mind (as described under the section "The Observer" on page 67) in order to assess whether the action she is considering asking about is newly demonstrated or has been executed multiple times in the past. She also checks the classroom context in order to understand the broader decision-making around the action in order to inform questions like, "What made you want to dismiss that hypothesis so early in your research?"

Like the Mentor, the Interviewer continually observes team's actions in MILA and responds when certain conditions are met. Unlike the Mentor, however, the Interviewer responds more directly to individual team actions. Every time the teams perform any action, the Interviewer is alerted to check her production rules for a matching rule. If one is triggered (typically involving the action itself as one of its rules), the Interviewer immediately alters her avatar to include the light bulb to let the team know it has a question to ask. Unlike the Mentor, the Interviewer does not overwrite past questions with future ones; the question will remain until the team answers the question or closes the software. Once the team answers the question, the Interviewer can optionally reply with some textual feedback, or she can simply thank the team for answering the question. Also unlike the Mentor, the Interviewer will ask certain questions more than once; other questions she will ask only once. Like the Mentor, though, the Interviewer does apply a threshold to prevent herself from asking questions too often, even though this means that sometimes she passes up the chance to act on some of her production rules.

Detailed Production Rules

Below are three examples of specific production rules used by the Interviewer of varying levels of complexity. Full rules can be found at dilab.gatech.edu/mila/

Table 12: One of the Interviewer's questions. This question can be asked multiple times as needed, although several dismissals in a row will not be considered individually due to the threshold.

Interviewer Rule #1	
Conditions	Actions
IF: The team has just dismissed a hypothesis; AND: It is relatively early in the lesson; OR: The dismissed model was relatively simple.	THEN: Ask, "What prompted you to dismiss that hypothesis so quickly?"; then, "Sometimes hypotheses don't go anywhere at all and can be dismissed pretty quickly, but remember to always have a reason to dismiss an earlier hypothesis!"

Table 13: Occasional events do occur where the Interviewer and the Mentor "step" on each other, such as when both are prepared to give feedback on hypothesis dismissals. In these instances, the Interviewer is capable of modifying the interval the Mentor uses to check its production rules, delaying the feedback and avoiding inundating the teams with too many interruptions.

Interviewer Rule #2	
Conditions	Actions
IF: The team has just dismissed a hypothesis; AND: It is relatively late in the lesson; OR: The dismissed model was relatively complex and justified with significant evidence.	THEN: Ask, "What prompted you to dismiss such a thorough model?"; then, "Dismissing hypotheses when you find contradictory evidence is crucial and represents a significant step forward for science. Scientists should never be ashamed to admit they are wrong in the face of contradictory evidence."

Table 14: A rule demonstrating the Interviewer's attempt to reason over the team's behavior and identify a relevant question to ask. If the team rapidly dismissed and reconsidered a model, it might be indicative of just testing out the function, player with the software, or humoring the Mentor's or a teacher's suggestion. Significant time between dismissal and reconsideration, however, suggests a thoughtful decision.

Interviewer Rule #3	
Conditions	Actions
IF: The team has just reconsidered a hypothesis; AND: Significant time has passed since the hypothesis was dismissed.	THEN: Ask, "What discovery prompted you to reconsider that hypothesis?"; then, "Sometimes you find evidence for an old, failed idea while working on a new one. Just like you shouldn't be afraid to dismiss ideas when they aren't going anywhere, you shouldn't be afraid to revisit them, either!"

On Metacognition

As mentioned previously, the primary role played by the Interviewer with regard to students' development of the desired metacognitive skills is to ask the students the questions that they ought to ask themselves. Self-reflection is a broader type of metacognition that is directly applicable in inquiry-driven modeling, and the Interviewer both encourages and demonstrates exactly this kind of self-reflection by prompting the team to reflect openly, modeling the behavior that they ought to internalize moving forward.

The Observer

Unlike the other tutors, the Observer has no visual manifestation and provides no feedback to the team. Rather, the Observer sits in the background, noting and tracking team behaviors and using them to modify an underlying profile of the team's ability. This profile, then, is available to the other tutors for use in their reasoning process. In the next chapter, the Control and Experimental groups in the deployment experiment are discussed; in the Control condition, teams do not see the tutors, but the Observer is still present in the background noting team behaviors.

Functional Role

When in a classroom or tutoring session, a teacher or human tutor is constantly noting elements of team's ability, tendencies, strengths, and weaknesses (Merrill, Reiser, Ranney, & Trafton 1992; Olney et al. 2012). These do not always prompt immediate feedback such as a correction or compliment, but contribute to an understanding of the student's ability that the teacher will use at a later time in service of another functional role. For example, a teacher (playing the functional role of the Mentor) might observe two students make the same error. The teacher may have previously noted that the first student has performed the skill correctly in the past while the second has made this same error repeatedly. Given that mental profile of the team's abilities, the teacher may vary their feedback, reminding the first student that they have done it correctly in the past and demonstrating the skill again more clearly for the second student.

Detailed Functional Information

Figure 20 below shows the input that the Observer receives and the output that it returns in response. The Observer operates on the raw log of actions that the team executes; each time an action is executed, the Observer reads it and decides whether to modify its profile of the team's ability. Unlike the other tutors, the Observer's output is entirely internal to the system. The Observer maintains three profiles of the team's ability: a profile of inquiry ability, a profile of modeling ability, and a profile of systems understanding.

Like the Interviewer, the Observer is alerted every time the team performs any action in the software; unlike the Interviewer, though, the Observer considers a longer history of the team's actions to attempt to identify the higher-level behavior. Upon being alerted that the team has performed some action, the Observer examines that action along with several of the other immediately preceding actions. The Observer then matches these actions against a set of action patterns it has, and if an action pattern is observed, the Observer modifies the underlying team profile. The Observer has separate team profiles for inquiry and modeling, reflecting related but distinct skills. For each skill, the Observer has low-level behaviors, such as Hypothesis Generation and Component Modification, and higher-level abstractions that summarize these lower-level behaviors. When the Observer notes the team performing an identified action pattern, it increments or decrements the portion of the profile corresponding to that pattern. It is important to note, however, that the Observer never considers a single performance of an action pattern to be indicative of the overall skill itself. Instead, the Observer demands a certain minimum threshold before considering that efficacy with a skill has been achieved. In this way, the Observer attempts to avoid some of the difficulties with error detection and just-in-time feedback characteristic of metacognitive tutoring systems (Roll et al. 2007); it requires a more persistent and repeated demonstration of a skill to accept that the skill has been mastered.

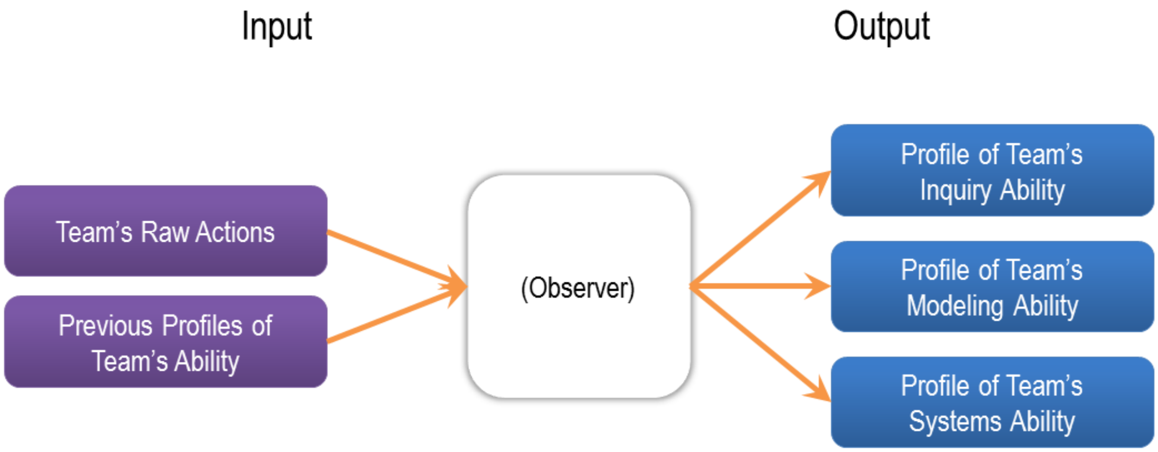

Input
Output
Team's Raw Actions
Previous Profiles of Team's Ability
(Observer)
Profile of Team's Inquiry Ability
Profile of Team's Modeling Ability
Profile of Team's Systems Ability

Figure 20: The inputs and outputs for the Observer. The Observer checks each action that the team takes and modifies its profile of the team's ability accordingly.

Detailed Production Rules

Below are three production rules used by the Observer of varying levels of complexity. As noted previously, incrementing the Observer's profile of a particular skill is not indicative of the Observer recognizing overall efficacy with that skill. Full rules can be found at dilab.gatech.edu/mila/

Table 15: The Observer modifies its profile of team behavior to reflect a potential increased ability to reason over and utilize multiple types of evidence, and a potential increased tendency to favor strongly-evidenced connections. The two low-level criteria on the right – Evidence Variety and Evidence Strength – then are united with other low-level behaviors regarding evidence to give an abstraction of the team's comfort with evidence as suggested in Guide Rule #2.

Observer Rule #1 Conditions	Actions
IF: The team has just added a new piece of evidence; AND: That new piece of evidence is of a type that has not previously been used; AND: That new piece of evidence creates an edge with an evidence score above 5.	THEN: Increment the Evidence Variety criteria of the Inquiry profile; AND: Increment the Evidence Strength criteria of the Inquiry profile.

Table 16: The Observer recognizes a more advanced explanation in the model and modifies a low-level criteria reflecting it. This criteria is then combined with other similar advanced modeling structures, such as Parallel Effects, Causal Flow, and Feedback Cycles, to give an overall profile of the team's modeling ability, as used in Critic Rule #2.

Observer Rule #2	
Conditions	Actions
IF: The team has just added a new connection to their model; AND: That new piece connection establishes a model demonstrating the complexity of parallel chains of causation.	THEN: Increment the Parallel Causation criteria of the Modeling profile.

Table 17: The Observer recognizes that the team is proposing a new hypothesis late in the curriculum and have already had some success modeling previous hypotheses, together suggesting a willingness to consider new ideas rather than stick to old ones. Assessing that as a positive behavior, the Observer increments its profile of the team's Hypothesis Generation behaviors more significantly than when a hypothesis is proposed on the first day (when such a hypothesis is necessary to begin).

Observer Rule #3	
Conditions	Actions
IF: The team has just added a new hypothesis to their project; AND: It is relatively late in the curriculum; AND: The team previously had one or more models meeting minimum level of thoroughness.	THEN: Significantly increment the Hypothesis Generation criteria of the Inquiry profile.

On Metacognition

All of the behaviors that the Observer attempts to identify are metacognitive in nature; the Observer uses the observable actions in the software as proxies for an underlying metacognitive trait. A crucial element of this, though, is that performance of the action a single time does not reflect mastery of that metacognitive skill; it could be misidentified by the Observer or it could be an anomaly in the team's behavior that they do not yet perform reliably. To account for this, the Observer demands repeated performance of the behaviors it uses as proxies for the underlying skills, assuming that repeated performance of a behavior indicative of the metacognitive skill implies greater probability that the skill is actually informing the behavior.

Summary of Tools

As described in this chapter, there are three primary tools at play in an intervention using MILA and MILA–T: a model of inquiry-driven modeling, an exploratory learning environment, and a metacognitive tutoring system. All three have been developed based on the project's ten-year history, the supporting literature, and our own experience with a pilot study in 2012. Although considerable more research can be performed on the model of inquiry-driven modeling and on MILA itself as an exploratory learning environment, the primary goal of the research presented here is to identify what, if any, learning gains are seen from interaction with the tutoring system compared to interaction with an identical project without the tutoring system. The experimental setup described in the next chapter, then, is a controlled experiment: given two groups of students with an identical curriculum and identical exploratory learning environment, to what extent do teams equipped with this metacognitive tutoring system outperform teams interacting without it?

CHAPTER 3

EXPERIMENTAL DESIGN & METHODS

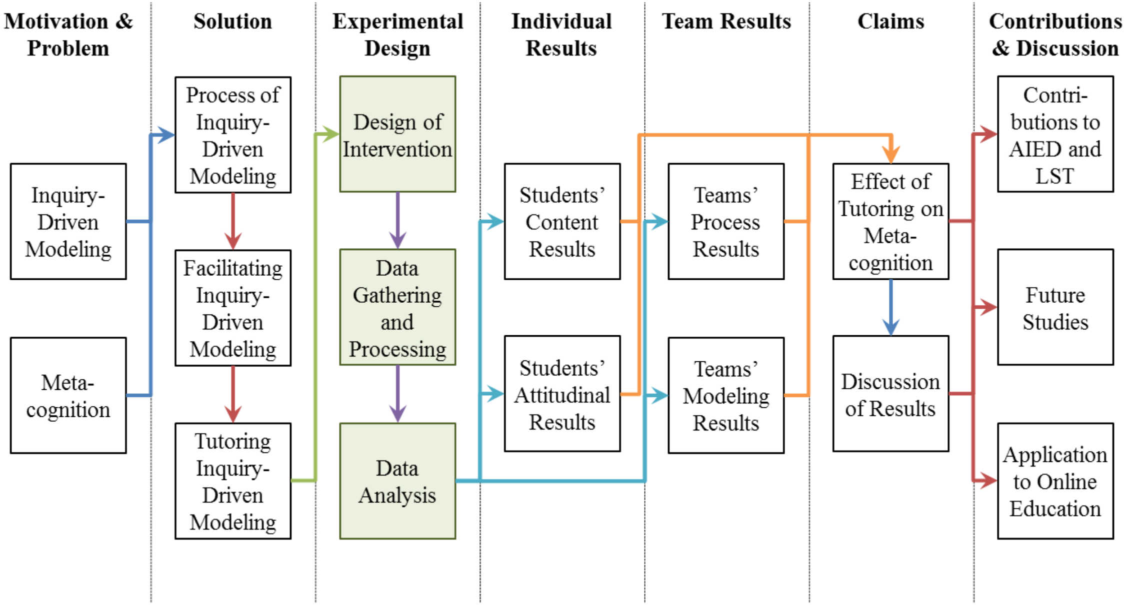

Motivation & Problem
Solution
Experimental Design
Individual Results
Team Results
Claims
Contributions & Discussion
Inquiry-Driven Modeling
Meta-cognition
Process of Inquiry-Driven Modeling
Facilitating Inquiry-Driven Modeling
Tutoring Inquiry-Driven Modeling
Design of Intervention
Data Gathering and Processing
Data Analysis
Students' Content Results
Students' Attitudinal Results
Teams' Process Results
Teams' Modeling Results
Effect of Tutoring on Meta-cognition
Discussion of Results
Contri-butions to AIED and LST
Future Studies
Application to Online Education

Figure 21: The third chapter of this dissertation outlines the experiment that was used to test the effectiveness of the previously-described tutoring system for teaching inquiry-driven modeling. It starts by describing the design of the experiment itself, then describes how the data were gathered, processed, and analyzed.

This chapter outlines the design of the experiment, including the arrangement of Control and Experimental groups (and other independent variables), the data gathering and processing procedures, and the analyses that have been run on those data. The results of these analyses are presented in full in Chapter 4 and 5, and the processed claims and conclusions are presented in Chapter 6.

Experimental Design

In this study, students participated in a nine-day curriculum on scientific modeling and inquiry. As part of this nine-day unit, students used the software system MILA for five class days. The curriculum was provided to the tutoring system MILA–T, and the tutors used this curriculum and its accompanying milestones to contextualize the expected level of student performance during the different days of the intervention. More detailed information about the curriculum in which students engaged can be found in Appendix C.

Students in the Experimental condition received MILA with the embedded tutoring system described previously for the first four days of the software interaction portion of the unit; students in the Control condition received MILA without the embedded tutoring system during these four days. On the fifth day, neither group received the embedded tutoring system. The first four days were spent addressing the problem of why thousands of fish suddenly died in Lake Clara Meer; henceforth, this is referred to as the Learning project because during this project, students in the Experimental group used MILA–T to improve their understanding (e.g. Learn) of inquiry-driven modeling. The fifth day was spent addressing the problem of why Atlanta has experienced significant temperature increases in the past several years; henceforth, this is referred to as the Transfer problem because it aims to test whether or not students in the Experimental group transferred any improvement in modeling ability to a new problem. Control groups received no tutor feedback in the Learning project, while Experimental groups did receive tutor feedback in the Learning project. Neither the Control nor Experimental groups received tutor feedback in the Transfer project.

This study was conducted with 276 students in seventh grade Life Science classes. Students were given the option of whether or not to consent to have their data used in the study; for those students who did not consent, their content tests, attitudinal surveys, and teams' models were not included in the analysis. 237 students consented to participate in the study. Given that the study was conducted in a normal classroom, limited control was available over assignment of subjects to Control and Experimental conditions. Two teachers administered the classroom during the study with limited intervention from the researchers. Each teacher taught five classes. Each individual class was labeled by the school as either Gifted, On-Level, or Inclusion, and each individual student was labeled by the school as either Gifted, High-Achieving, On-Level, or Special Needs. Gifted classrooms were comprised of Gifted and High-Achieving students; On-Level classes were comprised solely of On-Level students; Inclusion classes were comprised of both On-Level and Special Needs students. Given the presence of only seven high-achieving students and their participation in the Gifted classes, these students were classified as Gifted for analysis. Teacher A taught five classes: two Gifted classes, one On-Level class, and two Inclusion classes. Teacher B taught five classes: three On-Level classes and two Inclusion classes. Given the limited number of Special Needs students in the Inclusion classes and their individual identification for individual analysis, On-Level and Inclusion classes are grouped together for most team-level analysis.

Entire classes were assigned to Control and Experimental conditions in a pair-wise fashion based on scheduling restrictions. Each teacher was given three Experimental classes and two Control classes in order to facilitate comparisons within teachers. The Gifted classes of Teacher A were split between the Control and Experimental conditions, as were the Inclusion classes; the sole On-Level class for Teacher A was assigned to the Experimental group. The Inclusion classes of Teacher B were split between the Control and Experimental conditions, as were the On-Level classes; the third On-Level class was assigned to the Experimental condition. The chart below shows the assignment of classes to the Control and Experimental conditions.

Table 18: The assignment of classes to the Control and Experimental conditions.

	Control	**Experimental**
Teacher A	Class A (Inclusion) Class B (Gifted)	Class C (Inclusion) Class E (Gifted) Class D (On-Level)
Teacher B	Class F (On-Level) Class G (Inclusion)	Class I (On-Level) Class H (Inclusion) Class J (On-Level)

Individually at the beginning of the study, students took a content test and attitudinal survey. Afterward, students were assigned to teams of two or three for the duration of the study; team assignment was taken care of by the teachers without input from the researchers. Teachers later reported that they attempted to put students who had not consented to participate in the study in teams together to maximize the number of fully-consenting teams; this was not performed at the request of the researchers. 98 total teams were formed, with 84 teams fully comprised of students consenting to the study (three teams, however, did not complete the Transfer project due to absences). Data are available on the make-up of each team. At the conclusion of the study, students again took the content test and attitudinal survey individually.

A graphical calendar depicting the day-by-day activities of each class is shown below. Control and Experimental classes were exposed to the same lessons on the same days, with the only controlled variable being exposure to the tutoring system. On those days in which MILA was used, the teachers in each class began the class period with a short scripted lecture and a classroom activity. The short scripted lecture was the same for Control and Experimental classes. During this lecture, teachers explained to students how to interact with MILA, the nature of mechanism in MILA modeling, the importance, strength, and types of evidence that could be provided in MILA, and the overall criteria of a good model. During both classes as well, the teachers moved around the classroom giving teams feedback on the models they constructed and their overall modeling processes. No guidance was given to teachers to vary their behavior between Control and Experimental classes, although teachers later reflected that their behavior between Control and Experimental classes did differ due to the availability of MILA–T as an alternate source of feedback for students. Both Control and Experimental classes also participated in several shared activities outside MILA. In total, four lessons differed between the Control and Experimental classes.

Table 19: The lessons in which students participate during the intervention.

	Control Classes	Experimental Classes
05/06	Pre-Test and Pre-Survey, Introduction to the Lake Clara Meer Problem No MILA	
05/07	Lesson 1: Introduction to MILA With MILA, **without** tutors	Lesson 1: Introduction to MILA With MILA, **with** tutors
05/08	Lesson 2: Biological and Ecological Content Knowledge No MILA	
05/09	Lesson 3: Introduction to Mechanism With MILA, **without** tutors	Lesson 3: Introduction to Mechanism With MILA, **with** tutors
05/10	Lesson 4: Introduction to Simulations With MILA, **without** tutors	Lesson 4: Introduction to Simulations With MILA, **with** tutors
05/13	Lesson 5: Lake Water Quality Experiment No MILA	
05/14	Lesson 6: Free Modeling With MILA, **without** tutors	Lesson 6: Free Modeling With MILA, **with** tutors
05/15	Lesson 7: Atlanta Temperatures With MILA, **without** tutors	
05/16	Post-Test and Post-Survey, Q&A with Georgia Tech Researcher No MILA	

With the exception of interaction with the tutoring system, all elements of the projects were the same across the Control and Experimental classes. The same scripted introduction to MILA was performed by the teachers at the beginning of each class. On each of the following three days of interaction with MILA, students in both groups participated in a model walk where they looked at and evaluated MILA models assembled by others, followed by a discussion of the strengths and weaknesses of each model. Both groups discussed the different kinds of evidence, the increased value of certain kinds of evidence (such as Expert Information compared to Non-Expert

Information), and the criteria for evaluating a good model. In this way, the learning goal of inquiry-driven modeling was integrated into the curriculum of the unit and delivered to both the Control and the Experimental groups. While MILA–T aims to improve understanding of inquiry-driven modeling even further, understanding inquiry-driven modeling was nonetheless the explicit learning goal in the Control groups of the project as well.

Raw Data

Data generated over the course of this study fall into two groups: individual data and group data. Individual data are gleaned from individual students, whereas group data are generated by students working in teams of two or three where data points cannot be attributed to any individual student.

Individual

The study generated three forms of individual data: content test data, attitudinal survey data, and user experience data.

- **Content Test Data.** Students took a content test of 15 questions before and after the study. The content test examined biological content knowledge and understanding of the scientific process. Test questions were examined by an independent expert prior to usage. Pre-test results will be used to assess individual differences prior to engagement with the intervention.
- **Attitudinal Survey Data.** Students completed selections from two attitudinal inventories before and after the study: TOSRA (Test of Science Related Attitudes, Fraser 1981) and mATSI (Modified Attitudes Towards Science Inventory, Weinburgh & Steele 2000). Combined, they measure five constructs: attitude toward scientific inquiry, interest in careers in science, desire to do science, perception of the science teacher, and anxiety toward science. Pre-test results will

be used to assess individual differences prior to engagement with the intervention. These tests were chosen to examine multiple facets of students' attitudes toward science, in service of the dispositional learning goals outlined above and advocated by the literature (Roll et al. 2007). These five metrics are a subset of the twelve provided by the two inventories together and were chosen to limit the overall length of the survey to preserve student engagement. The metrics were chosen through consultation with an expert on science education.

- **User Experience Data.** At the conclusion of the study, students also completed a survey reflecting on their experience with the intervention. For the Experimental condition, this survey was comprised of three free-response questions about the tutors, one free-response question about the software as a whole, thirteen multiple choice questions about the tutors, and three multiple choice questions about the software as a whole. Control classes only received the one free-response question and three multiple choice questions about the software as a whole.

Team

During engagement with the software, all student activities were recorded in multiple ways. Additionally, final projects generated by students were recorded. These different means of tracking work out to several different individual types of raw team-level data:

- **Final Learning Project Models.** The first four days of the software interaction portion of the intervention, students completed models describing what they believed happened in Lake Clara Meer, as described previously. By the fourth day, students were asked to choose the final model in which they were most confident. This final model and the rejected models are recorded.
- **Final Transfer Project Models.** The fifth day of the software interaction portion of the intervention, students completed models describing why the city of Atlanta

has experienced record high temperatures the past several years, as described previously. Students were asked to submit one model of this phenomenon. This model and any rejected models are recorded.

- **Learning Project Interaction Log.** Within the software, every individual interaction that students complete is recorded. Thus, a comprehensive log of the entire model construction and revision process that students complete while engaging in the software was recorded.
- **Transfer Project Interaction Log.** Similarly, a revision log of students' process of constructing a model of temperature change in Atlanta was recorded.
- **Software Usage Statistics.** In addition to the revision log, the software also tracks more basically the amount of time that students spend engaged with the software. This gives a view as to the amount of exposure the different teams receive to the treatment condition.
- **Tutor Interaction Statistics.** In addition to the revision log, the software also tracks all interactions between the students and the tutors. This includes the times when the Mentor and Interviewer had available feedback that the team did not consult, and the profile that the Observer developed of teams over time.

Data Cleaning & Processing

Prior to analysis, data drawn from the experiment were cleaned and processed. In the case of the individual data, this process was relatively simple, consisting of removing the data for non-consenting students and evaluating the content test and attitudinal surveys. For the team-level data, more significant cleaning and processing was required.

Individual

First, content tests and attitudinal surveys from non-consenting students were removed from the study's data pool. The remaining 237 content tests and attitudinal surveys were then paired within students and graded. The pre-test score, the post-test score, and the change between the two were recorded for the content test. Similarly, the pre-survey score, the post-survey score, and the change between the two were recorded for each of the five attitudinal surveys. The result of this cleaning process is a set of three scores for each of the six metrics for each student: the score before the intervention, the score after the intervention, and the change between the two.

Team

More significant data cleaning and processing was required for team-level data given that the data are not initially numeric. This process resulted in three types of data:

- **Final Model Summaries.** During the intervention, each team produced two finals models: a final model of the Learning project (the Lake Clara Meer problem) and a final model of the Transfer project (the Atlanta Temperature problem). Analysis of these models identified qualitative differences in the models generated with tutor feedback and those generated without tutor feedback. Models were scored according to the following criteria: model complexity; total model strength; average model strength; average evidence strength; and total evidence. Model complexity sums the components and connection in the model to summarize its size and interconnectedness. The four evidence measurements identify different types of evidential strengths behind a model: how much evidence is given for the model as a whole (total model strength), how much evidence is given for each individual claim within the model (average model strength), how strong each individual piece of evidence is (average evidence strength), and how many total pieces of evidence are given in the model (total evidence).
- **Model Construction Statistics.** In order to assess the model revision process in which students engaged, the low-level software interactions were coded into higher-level activities corresponding to the inquiry-driven modeling process

described previously. Based on these activities, Markov chains were calculated to summarize the transitions between different activities in the inquiry-driven modeling process and compare those transitions to the ideal execution of the process.

- **Tutor Interaction Statistics.** The software logs each individual interaction that teams have with the tutoring system, along with timestamps indicating when the interaction ceased or how long the interaction lasted. These raw data were processed to give a clean account of the frequency with which each team consulted the tutors, which tutor they consulted, and the amount of time paid to the feedback itself as suggested by the timestamps revealing the time in which the feedback was visible.

At the conclusion of this process, the clean team-level data provided numeric summarizations of students' models, students' model revision process, and students' interactions with the tutoring system.

Variables

The data cleaning and processing process described above generated numerical data on a variety of elements of the intervention. These data, in turn, becomes the input and output variables of the following study. Again, these are split between individual and team variables.

Individual

At the individual level, the analysis focuses on the effect of exposure to the Experimental condition on students' performance and responses on the pre- and post-test and -survey.

Input Variables

The primary input variable is exposure to the experimental condition. Altogether, 99 consenting students were exposed to the Control condition and 138 consenting students were exposed to the Experimental condition. In addition to the Control and Experimental conditions, there are two other variables to be analyzed as controlled input variables:

- **Teacher.** Based on casual observation, significant differences were seen in the workings of the two teachers' classrooms. It is possible, and in fact likely, that there exist differences in performance based on differing teacher behavior.
- **Student Level.** Individual students were given different identifiers by the school: Gifted, High-Achieving, On-Level, and Special Needs. Results from engagement with the intervention may differ based on students' prior success and ability, as summarized in these labels.

Output Variables

At the individual level, there are three categories of output variables. These variables will be further sub-divided for analysis, but retain this categorical structure.

- **Change in Content Test Results.** Students' change in content test scores is analyzed as a product of the intervention. In addition to the change, the analysis also examines whether there are any initial or final differences in content knowledge between the Control and Experimental groups.
- **Change in Attitudinal Survey Results.** Students' change in attitudinal survey scores is calculated for each of the individual inventories of the TOSRA and mATSI. In addition to the change, the analysis also examines whether there are any initial or final differences in attitudes between the Control and Experimental groups.

- **User Experience Survey Results.** This variable summarizes students' self-reported experience with the software and tutoring system. This is used to examine whether there exists a difference in opinion toward the software based on exposure to the Experimental condition.

Team

At the team level, the analysis focuses on the effect of exposure to the Experimental condition on students' models and inquiry-driven modeling process for both the Learning project and the Transfer project.

Input Variables

As with the individual analysis, the primary input variable in the team analysis is the Control or Experimental condition. Altogether, 84 fully-consenting teams used the software, with 50 in the Experimental condition and 34 in the Control condition. The primary controlled variables for the team-level analysis mirror those from the individual data:

- **Teacher.** The two different teachers in the study had significantly different approaches to guiding teams' work, and thus projects between the teachers are anticipated to have significant differences.
- **Class Level.** The teachers' approaches to the different classes differed notably, especially in Teacher A's Gifted classes. Information on class labels will be included to examine differences between models and processes of teams in the Gifted classes and those in On-Level classes.

Output Variables

There are three primary categories of output variables available at the team level, corresponding to the data cleaning and processing procedures outlined above. All three of these variable categories exist for both the Lake Model and the Temperature Model.

- **Final Model Summaries.** Numeric summaries of model complexity and evidential strength are examined to test for differences in modeling behavior and quality between the conditions. These numeric summaries focus on the depth and complexity of teams' models and the depth, breadth, and strength of evidence used to justify teams' models. As part of the analysis, the evidence provided in these models is coded for usefulness and proper categorization, although this may be considered a portion of the cleaning process as well.
- **Model Construction Statistics.** In addition to testing the final models, the process of modeling is explicitly examined by looking at the Markov chains. These Markov chains are mapped to the inquiry-driven modeling process and examined first for improvements and second for overall differences (Kemeny 1960). Markov chains are chosen here rather than Markov models in order to test for compatibility with the prior model of inquiry-driven modeling rather than to determine a model of teams' execution of the skill from scratch (Ghahramani 2001).

Analysis

As with the data presented above, analysis is segmented between individual analysis and team analysis. Individual analysis addresses the first two high-level hypotheses, and team analysis addresses the second two high-level hypotheses (as well as the two sub-hypotheses under each).

Individual

The individual data are tested using a multivariate analysis of variance. First, a multivariate analysis of variance is used to establish whether or not differences existed in content knowledge or attitudes between the Control and Experimental groups prior to the study (I.1); if such differences exist, differences between the Control and Experimental groups could be attributed to these initial differences rather than the study. This initial analysis also identifies whether there exist any interactions between the other input variables (Teacher and Student Level) and the six metrics, although the presence of these interactions would not pose a problem for the study. After this preliminary analysis, the individual analysis addresses changes across each of the six metrics. This begins with a multivariate analysis of variance of change in all six variables together in order to limit repeated testing and reduce the risk of detecting false positives. This analysis considers the three input variables (Condition, Teacher, and Student Level) and changes in each of the six output variables (I.2). If this study detects interactions with any of these independent variables, a series of subsequent univariate analyses is run to test for interactions between each input variable and each output variable (I.2.A-F). Finally, for those metrics that demonstrate differences between the Control and Experimental groups, a follow-up t-test is used to identify whether students within either Condition changed in a statistically significant way (I.3), and whether the changes led to statistically significant differences in post-test scores between Control and Experimental groups (I.4). Hypothesis #1 (Explicit Understanding) is addressed by I.1, I.2, I.2.A, I.3, and I.4. Hypothesis #2 (Dispositional Framing) is addressed by I.1, I.2, I.2.B, I.2.C, I.2.D, I.2.E, I.2.F, I.3, and I.4. The overlap between I.1, I.3, and I.4 is due to simultaneously analyzing content and attitudinal data to minimize risk of type I error. Table 20 summarizes these studies.

Table 20: Studies examining the individual-level data.

Identifier	Study Description
I.1	Multivariate analysis of variance testing whether there existed prior differences based on Condition, Teacher, or Student Level.
I.2	Multivariate analysis of variance testing whether change in individual variables interacted with Condition, Teacher, or Student Level.
I.2.A	Univariate analysis of variance testing for an interaction between change in content score and Condition, Teacher, or Student Level.
I.2.B	Univariate analysis of variance testing for an interaction between change in Attitude toward Scientific Inquiry score and Condition, Teacher, or Student Level.
I.2.C	Univariate analysis of variance testing for an interaction between change in Career Interest in Science score and Condition, Teacher, or Student Level.
I.2.D	Univariate analysis of variance testing for an interaction between change in Anxiety toward Science score and Condition, Teacher, or Student Level.
I.2.E	Univariate analysis of variance testing for an interaction between change in Desire to Do Science score and Condition, Teacher, or Student Level.
I.2.F	Univariate analysis of variance testing for an interaction between change in Perception of the Science Teacher score and Condition, Teacher, or Student Level.
I.3	Univariate analysis of differences within groups between pre-test and post-test scores for metrics showing relevant differences in I.2.
I.4	Univariate analysis of differences across groups in post-test for metrics showing relevant differences in I.2.

Team

For team-level data, separate analyses are used to address the process data and the final models that teams of students produce. First, for the process data, Markov chains are created connecting the individual actions teams take within the model construction process to the underlying activities in the inquiry-driven modeling process. Markov chains were chosen instead of hidden Markov models because of the desire to map and compare teams' behaviors to a preexisting model of the target process or skill rather than develop a grounded description of teams' behaviors (Ghahramani 2001). The Markov chains derived from patterns of teams' engagement with the model construction process are then aligned to the inquiry-driven modeling process provided in Figure 4 on page 25. A significant limitation of analysis based on Markov chains is the assumption that the likelihood of a subsequent state is based solely on the current state rather than anything about the history of state transitions preceding the current state (Lopez, Hermanns, & Katoen 2001); however, for this initial analysis and for comparison to the desired process presented previously, this assumption is not unreasonable.

Five such chains are identified: one for each combination of Condition and Project (Control group Learning project, Control group Transfer project, Experimental group Learning project, Experimental group Transfer project), and an additional Markov chain including tutor interactions during the Experimental group's Learning project. Three studies are run on these Markov chains. First, based on identifiable improvements in Markov chains identified prior to processing these data, the Markov chains for the Experimental group are tested against the Markov chains for the Control group for discernible improvements (T.1L, T.1T). For this, a repeated Z-test will be conducted to test for differences between the Control and Experimental groups along several metrics. A repeated Z-test is chosen in order to target specific improvements within the broader Markov chains. A Bernoulli trial is used to correct for the increased likelihood of false positives.

After testing for targeted improvement, a χ^2 analysis is run on the overall results to test for general differences to ascertain whether or not the presence of the tutoring system actually changed the inquiry-driven modeling process within either project (T.2L, T.2T). A χ^2 analysis identifies whether sets of proportions are different between different groups, and thus this analysis will identify whether or not Control and Experimental groups are allocating their efforts differently among the phases of the process. A χ^2 will not, however, identify whether or not one group actually executes all tasks more often than the other if the proportions are the same; initial preliminary analysis suggested analysis along this path was not worth pursuing. Finally, a case study analysis is performed on the fifth Markov chain to ascertain where in the inquiry-driven modeling process tutor feedback was invoked and what activities typically followed tutor feedback (T.3). These five analyses address Hypothesis #3 (Procedural Execution). Hypothesis #3A is addressed by T.1L and T.2L, while Hypothesis #3B is addressed by T.1T and T.2T. T.3 does not provide evidence directly for or against any part of Hypothesis #3, but provides a more complete picture of the way in which the tutoring system was used by Experimental teams.

Analysis of the final models and explanations teams produce is analogous to the studies performed on the individual data. First, a multivariate analysis of variance is performed with three input variables (Condition, Teacher, Class Level) and five output variables (the five measures of model quality) within the Learning project (T.4L) and the Transfer project (T.4T). If interactions are detected during that multivariate analysis, a follow-up univariate analysis of variance is performed on each individual output variable to test for interactions with each individual input variable on the Learning project (T.4L.A-E) and the Transfer project (T.4T.A-E) separately. After those initial analyses, the individual pieces of evidence that teams wrote are coded as acceptable or unacceptable (through a more complex coding scheme). Three rounds of coding are run by one coder with three weeks between subsequent sessions, with an intermediate study assessing the level of intrarater reliability. If sufficient intrarater reliability is established, a χ^2 analysis is conducted testing for differences in teams' performance as part of the evidence coding exercise, primarily testing to see whether Control or Experimental teams demonstrate a superior use of evidence within the Learning project (T.6L) or the Transfer project (T.6T). Finally, if sufficient intrarater reliability is established in T.5, the initial analyses are executed again with the models reinterpreted according to the results of the evidence coding exercise. Unacceptable pieces of evidence are thrown out and miscategorized pieces of evidence are scored under the correct category to recalculate the four evidence-based metrics from the initial analysis. The multivariate analysis of variance from the original analysis is run again within the Learning (T.7L) and Transfer (G.7T) projects, and if interactions are observed, the follow-up univariate studies from the original analysis is run (T.7L.A-E, T.7T.A-E). These analyses, T.4 through T.7, address Hypothesis #4 (Models and Explanations). Hypothesis #4A is addressed by T.4L, T.6L, and T.7L, while Hypothesis #4B is addressed by T.4T, T.6T, and T.7T. T.5 provides the information necessary to perform the other analyses described here. Table 21 summarizes these studies.

Table 21: Studies examining the group-level data.

Identifier	Study Description
T.1L	Repeated Z-test testing for desirable differences between the Control and Experimental groups' Markov chains during the Learning project.
T.1T	Repeated Z-test testing for desirable differences between the Control and Experimental groups' Markov chains during the Transfer project.
T.2L	χ^2 analysis testing for difference between Control and Experimental groups' Markov chains during the Learning project.
T.2T	χ^2 analysis testing for difference between Control and Experimental groups' Markov chains during the Transfer project.
T.3	Case study analysis on Markov chain for Experimental group's Learning project examining the position of the tutoring system in the process.
T.4L	Multivariate analysis of variance testing for an interaction between any independent variable (Condition, Teacher, Class Level) and any dependent variable (the five metrics for model evaluation) during the Learning project.

T.4L.A-E	Univariate analysis of variance for interaction between each input variable and each individual output variable during the Learning project.
T.4T	Multivariate analysis of variance testing for an interaction between any independent variable (Condition, Teacher, Class Level) and any dependent variable (the five metrics for model evaluation) during the Transfer project.
T.4T.A-E	Univariate analysis of variance for interaction between each input variable and each individual output variable during the Transfer project.
T.5	Intrarater reliability analysis on three rounds of evidence coding.
T.6	χ^2 analysis of difference in performance between the Control and Experimental groups in the evidence coding exercise.
T.6L	χ^2 analysis of difference in performance between the Control and Experimental groups in the evidence coding exercise on the Learning project.
T.6T	χ^2 analysis of difference in performance between the Control and Experimental groups in the evidence coding exercise on the Transfer project.
T.7L	Multivariate analysis of variance testing for an interaction between any independent variable (Condition, Teacher, Class Level) and any dependent variable (the five metrics for model evaluation) during the Learning project after the evidence coding process.
T.7L.A-E	Univariate analysis of variance for interaction between each input variable and each individual output variable during the Learning project after the evidence coding process.
T.7T	Multivariate analysis of variance testing for an interaction between any independent variable (Condition, Teacher, Class Level) and any dependent variable (the five metrics for model evaluation) during the Transfer project after the evidence coding process.
T.7T.A-E	Univariate analysis of variance for interaction between each input variable and each individual output variable during the Transfer project after the evidence coding process.

CHAPTER 4

INDIVIDUAL RESULTS

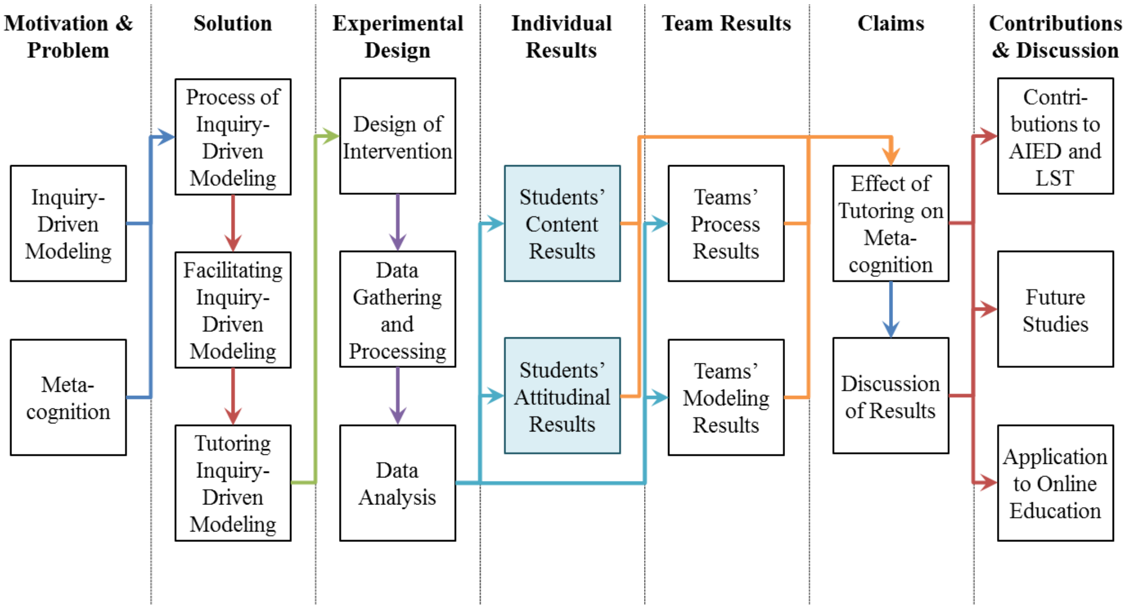

Motivation & Problem
Solution
Experimental Design
Individual Results
Team Results
Claims
Contributions & Discussion
Inquiry-Driven Modeling
Meta-cognition
Process of Inquiry-Driven Modeling
Facilitating Inquiry-Driven Modeling
Tutoring Inquiry-Driven Modeling
Design of Intervention
Data Gathering and Processing
Data Analysis
Students' Content Results
Students' Attitudinal Results
Teams' Process Results
Teams' Modeling Results
Effect of Tutoring on Meta-cognition
Discussion of Results
Contri-butions to AIED and LST
Future Studies
Application to Online Education

Figure 22: The fourth chapter of this dissertation covers the results of the individual analyses on students' declarative understanding of and dispositional orientation towards inquiry-driven modeling.

In line with the hypotheses presented previously, the analysis of these data aims to identify the changes in students' understanding of, dispositions toward, process of, and creations through inquiry-driven modeling. This chapter will start by examining the individual components of this research: individual students' understanding of and dispositions toward inquiry-driven modeling. First, it will first look at the state of the students prior to the intervention. Second, it will examine the results of the first hypothesis, addressing students' understanding of inquiry-driven modeling before and after the intervention. Third, it will examine the results of the second hypothesis, addressing students' dispositions towards science, scientific inquiry, and careers in science. Finally, it will look at the results of a selection of additional sources of data that do not align directly with these hypotheses, but nonetheless contribute to an overall view of the utility of the tutoring system.

Data Structure

Throughout this chapter, several tables will be presented that summarize the numeric results from the Control and Experimental groups. For the sake of completeness, these tables present the mean, standard deviation, and number of subjects within every significant group, from the Control and Experimental groups as a whole to narrower distinctions such as Teacher A's On-Level Control group students. Because of the quantity of data presented in these tables, they may be rather complex. In order to help understand these tables, the general structure of these data are presented below.

Table 22: The generic structure of the tables presented in this chapter. Each cell of these tables will provide three numbers: the mean on top, the standard deviation in parentheses in the middle, and the number of subjects at the bottom.

General Individual Data Table Structure

<table>
<tr><td colspan="3">Teacher A</td><td colspan="3">Teacher B</td><td colspan="3">Teacher A</td><td colspan="3">Teacher B</td></tr>
<tr><td colspan="3">Teacher A's Control group students</td><td colspan="3">Teacher B's Control group students</td><td colspan="3">Teacher A's Experimental group students</td><td colspan="3">Teacher B's Experimental group students</td></tr>
<tr><td colspan="6">Control</td><td colspan="6">Experimental</td></tr>
<tr><td colspan="6">All Control group students</td><td colspan="6">All Experimental group students</td></tr>
<tr><td colspan="2">On-Level</td><td colspan="2">Gifted</td><td colspan="2">Sp. Needs</td><td colspan="2">On-Level</td><td colspan="2">Gifted</td><td colspan="2">Sp. Needs</td></tr>
<tr><td colspan="2">On-Level Control group students</td><td colspan="2">Gifted Control group students</td><td colspan="2">Special Needs Control group students</td><td colspan="2">On-Level Experimental group students</td><td colspan="2">Gifted Experimental group students</td><td colspan="2">Special Needs Control Experimental students</td></tr>
<tr><td>A</td><td>B</td><td>A</td><td>B</td><td>A</td><td>B</td><td>A</td><td>B</td><td>A</td><td>B</td><td>A</td><td>B</td></tr>
<tr><td>Teacher A's On-Level Control group students</td><td>Teacher B's On-Level Control group students</td><td>Teacher A's Gifted Control group students</td><td>Teacher B's Gifted Control group students</td><td>Teacher A's Sp. Needs Control group students</td><td>Teacher B's Sp. Needs Control group students</td><td>Teacher A's On-Level Exper-imental group students</td><td>Teacher B's On-Level Exper-imental group students</td><td>Teacher A's Gifted Exper-imental group students</td><td>Teacher B's Gifted Exper-imental group students</td><td>Teacher A's Sp. Needs Exper-imental group students</td><td>Teacher B's Sp. Needs Exper-imental group students</td></tr>
</table>

Each cell in the tables that follow provides the subscores for the group of students indicated in Table 22. Within each cell are three numbers: the number at the top is the mean for that group of students; the number in parentheses in the middle is the standard deviation for that group of students; and the number at the bottom, preceded by "*n* =", is the number of subjects in that group of students. Teacher B has no Gifted students, and as such the 'Teacher B's Gifted' student cells are empty. The table is structured to allow as many comparisons as possible under the general structure of the experiment. We can, for instance, directly compare the Control and Experimental students as a whole by comparing the 'All Control group students' and 'All Experimental group students' cells, or we may compare only On-Level students by comparing the 'On-Level Control group students' and 'On-Level Experimental group students'. In reading the tables that follow, please consult Table 22 to ensure understanding of the structure of the data presented.

Preliminary Analysis

Prior to conducting the analysis of the outcomes of the intervention, two analyses were necessary to assert the value and validity of the following analyses. First, it is necessary to assert that there exists no difference between the Control and Experimental groups prior to the intervention. Second, it is useful to understand differences based on Teacher or Student Level prior to the intervention.

Equivalence of Means Prior to Intervention

As stated in the previous chapter, entire classes were assigned to either the Control or the Experimental condition in order to prevent students from being aware of aid being given to other teams in their classes. Classes were also divided up in order to ensure an even distribution of Inclusion, On-Level, and Gifted classes between the two conditions, leading to the assignments noted in Table 18. Given this basis for the assignment to conditions, it is possible that the two groups started the experiment out unequal according to individual metrics that could be assessed prior to conducting the analysis. Thus, the first analysis examined potential differences between Control and Experimental students prior to the intervention (I.1). This analysis looks at pre-test scores for students on the content test, as well as all five attitudinal constructs.

Table 23 presents the results of this initial analysis. In order to minimize the risk of Type I error, all independent and dependent variables are analyzed in a single multivariate ANOVA. Thus, this table will be consulted further in the following section for the effects of the Teacher and the students' Levels on pre-test scores. Similarly, in the following sections, a single multivariate ANOVA will examine the influence of these factors on the change in both students' content and students' attitudinal scores.

Table 23: Multivariate ANOVA examining the influence of Condition, Teacher, and Level on students' pre-test scores. The table suggests that there are influences of the Teacher and class Level on students' pre-test scores, but no influence of Condition.

```
Multivariate ANOVA :
                         Df   Pillai approx F num Df den Df    Pr(>F)
Condition                 1 0.024370   0.9242      6    222   0.47830
Teacher                   1 0.058884   2.3150      6    222   0.03452
Level                     2 0.253933   5.4052     12    446 1.347e-08
Condition:Teacher         1 0.023896   0.9058      6    222   0.49150
Condition:Level           2 0.053265   1.0169     12    446   0.43186
Teacher:Level             1 0.012032   0.4506      6    222   0.84410
Condition:Teacher:Level   1 0.022925   0.8681      6    222   0.51908
Residuals               227

 Response Content.B :
                         Df  Sum Sq Mean Sq F value    Pr(>F)
Condition                 1   10.09  10.087  1.2210  0.270330
Teacher                   1   70.59  70.594  8.5452  0.003815
Level                     2  300.69 150.343 18.1986 4.673e-08
Condition:Teacher         1    5.23   5.233  0.6335  0.426914
Condition:Level           2   39.87  19.936  2.4132  0.091822
Teacher:Level             1    4.24   4.243  0.5135  0.474344
Condition:Teacher:Level   1    8.76   8.762  1.0606  0.304164
Residuals               227 1875.31   8.261

 Response AtSI.B :
                         Df  Sum Sq Mean Sq F value   Pr(>F)
Condition                 1    14.9   14.87  0.3040 0.581915
Teacher                   1    23.4   23.37  0.4778 0.490128
Level                     2   677.5  338.75  6.9256 0.001204
Condition:Teacher         1    14.7   14.65  0.2995 0.584723
Condition:Level           2    40.4   20.20  0.4130 0.662168
Teacher:Level             1     0.4    0.40  0.0081 0.928334
Condition:Teacher:Level   1     1.4    1.41  0.0288 0.865368
Residuals               227 11103.3   48.91
```

```
 Response CIiS.B :
                         Df  Sum Sq Mean Sq F value  Pr(>F)
Condition                 1    10.7  10.700  0.1547 0.69449
Teacher                   1     7.6   7.590  0.1097 0.74079
Level                     2   372.5 186.242  2.6920 0.06991
Condition:Teacher         1    11.4  11.441  0.1654 0.68464
Condition:Level           2   340.3 170.130  2.4591 0.08779
Teacher:Level             1    11.6  11.636  0.1682 0.68211
Condition:Teacher:Level   1    51.0  51.004  0.7372 0.39146
Residuals               227 15704.9  69.184

 Response AtS.B :
                         Df Sum Sq Mean Sq F value   Pr(>F)
Condition                 1    1.1   1.114  0.0754 0.783861
Teacher                   1    4.7   4.709  0.3189 0.572853
Level                     2  145.0  72.480  4.9078 0.008191
Condition:Teacher         1   10.3  10.257  0.6945 0.405515
Condition:Level           2   65.4  32.677  2.2127 0.111765
Teacher:Level             1    1.0   0.959  0.0650 0.799040
Condition:Teacher:Level   1   35.6  35.609  2.4112 0.121866
Residuals               227 3352.4  14.768

 Response PofTS.B :
                         Df  Sum Sq Mean Sq F value Pr(>F)
Condition                 1    1.33  1.3349  0.2742 0.6010
Teacher                   1    6.88  6.8847  1.4144 0.2356
Level                     2    5.22  2.6113  0.5365 0.5856
Condition:Teacher         1    1.33  1.3295  0.2731 0.6018
Condition:Level           2   14.04  7.0180  1.4417 0.2387
Teacher:Level             1    0.02  0.0164  0.0034 0.9538
Condition:Teacher:Level   1    0.01  0.0095  0.0020 0.9648
Residuals               227 1104.97  4.8677

 Response DtDS.B :
                         Df Sum Sq Mean Sq F value Pr(>F)
Condition                 1   61.2  61.244  1.9920 0.1595
Teacher                   1    3.3   3.274  0.1065 0.7445
Level                     2   20.5  10.264  0.3338 0.7165
Condition:Teacher         1   54.1  54.138  1.7608 0.1859
Condition:Level           2   76.3  38.149  1.2408 0.2911
Teacher:Level             1   17.9  17.930  0.5832 0.4459
Condition:Teacher:Level   1    2.3   2.284  0.0743 0.7854
Residuals               227 6979.3  30.746
```

Table 23 provides the output of a multivariate ANOVA on students' pre-test results on the content test and each of the five attitudinal constructs. The initial multivariate ANOVA, shown first, demonstrates that **no significant difference existed between the students in the Control and Experimental groups prior to the intervention**, as measured by these six metrics. Further ANOVA on each of the six individual constructs reveals no statistically significant difference according to any of these constructs between the Control and Experimental groups. Given this analysis, it is reasonable to assume that the Control and Experimental groups began the intervention on approximately equal footing in their content knowledge and attitude towards science based on these six metrics. Students in one group did not exhibit superior content knowledge or attitudinal scores than students in the other group

Comparison of Groups Prior to Intervention

Although there were no significant differences on the pre-test results between the Control and Experimental conditions, there did exist significant differences based on the two other independent variables, Teacher and Level. Table 23 further examines the relationship between these two variables and the results of the pre-test. First, the multivariate ANOVA reveals a relationship between Teacher and pre-test performance (F = 2.32, $p < 0.05$), and between student Level (Gifted, On-Level, Special Needs) and pre-test performance (F = 5.41, $p < 0.001$). Further ANOVA of the specific individual reveals several differences. The Teacher determined statistically significant differences in students' pre-test content scores (F = 8.55, $p < 0.01$), with students of Teacher A outperforming students of Teacher B by an average of roughly one point. Similarly, statistically significant effects of the student Level were observed on students' content test scores (F = 18.20, $p < 0.001$) as Gifted students outperformed On-Level students, who outperformed Special Needs students.

No effect of the Teacher was observed on any attitudinal constructs. However, significant effects of class Level were observed on attitudinal metrics: attitude toward scientific inquiry (F = 6.93, $p < 0.01$), career interest in science (F = 2.69, $p < 0.1$), and anxiety toward science (F = 4.91, $p < 0.01$). Interestingly, Special Needs students expressed the most positive attitudes toward scientific inquiry, while Gifted students reported more positive attitudes than On-Level students. Gifted students reported the highest interest in a career in science, followed by On-Level students, followed by

Special Needs students. Gifted students also reported the least anxiety toward science, followed by On-Level students, followed by Special Needs students. No significant differences were observed in perception of the science teacher or desire to do science between either the Teachers or the different student Levels.

To a large extent, the results of this pre-test analysis are not surprising. Students assigned to the Gifted classes due to prior positive performance in science demonstrate higher mastery of scientific content knowledge and generally more positive attitudes toward science. Given that Gifted, On-Level, and Special Needs students are evenly balanced between Control and Experimental groups (that is, the ratio of students in the Control condition to students in the Experimental condition is the same across all three groups and both teachers), later analysis will be able to account for differences based on Level and Teacher in addition to Condition.

Table 24: Initial scores for each content and attitudinal metric for students in each group and subgroup. The top number in each cell is the mean for that group; the number in parentheses is the standard deviation within that group; the bottom number is the number of samples within that subgroup. See Table 22 for further information about the groups of students summarized in each cell.

Initial Content Scores

Teacher A			Teacher B			Teacher A			Teacher B		
10.22 (3.31) *n* = 45			8.37 (3.32) *n* = 54			9.93 (2.86) *n* = 62			9.38 (3.05) *n* = 76		
Control						Experimental					
9.21 (3.35) *n* = 99						9.63 (2.97) *n* = 138					
On-Level		Gifted		Sp. Needs		On-Level		Gifted		Sp. Needs	
8.93 (3.01) *n* = 71		12.44 (1.54) *n* = 18		5.40 (3.13) *n* = 10		9.46 (2.99) *n* = 113		11.50 (1.59) *n* = 16		8.44 (3.40) *n* = 9	
A	B	A	B	A	B	A	B	A	B	A	B
9.48 (2.64) *n* = 23	8.67 (3.16) *n* = 48	12.44 (1.54) *n* = 18	-- -- *n* = 0	4.50 (1.29) *n* = 4	6.00 (3.95) *n* = 6	9.55 (2.90) *n* = 38	9.41 (3.05) *n* = 75	11.50 (1.59) *n* = 16	-- -- *n* = 0	8.62 (3.58) *n* = 8	7.00 -- *n* = 1

Initial Attitude toward Scientific Inquiry Scores

Teacher A			Teacher B			Teacher A			Teacher B		
20.93 (6.86) *n* = 45			21.39 (7.48) *n* = 54			21.45 (7.86) *n* = 62			20.04 (6.30) *n* = 76		
Control						Experimental					
21.18 (7.17) *n* = 99						20.67 (7.05) *n* = 138					
On-Level		Gifted		Sp. Needs		On-Level		Gifted		Sp. Needs	
20.35 (7.14) *n* = 71		20.78 (7.24) *n* = 18		27.80 (3.19) *n* = 10		20.33 (6.38) *n* = 113		20.31 (8.19) *n* = 16		25.56 (11.24) *n* = 9	
A	B	A	B	A	B	A	B	A	B	A	B
20.00 (6.63) *n* = 23	20.52 (7.43) *n* = 48	20.78 (7.24) *n* = 18	-- -- *n* = 0	27.00 (3.74) *n* = 4	28.33 (3.01) *n* = 6	21.03 (6.55) *n* = 38	19.99 (6.32) *n* = 75	20.31 (8.20) *n* = 16	-- -- *n* = 0	25.75 (11.99) *n* = 8	24.00 -- *n* = 1

Initial Career Interest in Science Scores

Teacher A			Teacher B			Teacher A			Teacher B		
27.73 (7.11) *n* = 45			27.44 (9.11) *n* = 54			27.37 (8.87) *n* = 62			26.96 (8.22) *n* = 76		
Control						Experimental					
27.58 (8.22) *n* = 99						27.14 (8.49) *n* = 138					
On-Level		Gifted		Sp. Needs		On-Level		Gifted		Sp. Needs	
27.77 (8.21) *n* = 71		30.39 (7.99) *n* = 18		21.10 (5.30) *n* = 10		26.70 (8.26) *n* = 113		29.44 (7.81) *n* = 16		28.67 (12.24) *n* = 9	
A	B	A	B	A	B	A	B	A	B	A	B
27.00 (5.71) *n* = 23	28.15 (9.21) *n* = 48	30.39 (7.99) *n* = 18	-- -- *n* = 0	20.00 (3.74) *n* = 4	21.83 (6.37) *n* = 6	26.05 (8.33) *n* = 38	27.03 (8.26) *n* = 75	29.44 (7.81) *n* = 16	-- -- *n* = 0	29.50 (12.81) *n* = 8	22.00 -- *n* = 1

Initial Anxiety toward Science Scores

Teacher A			Teacher B			Teacher A			Teacher B		
10.71 (3.91) *n* = 45			11.19 (4.12) *n* = 54			11.56 (4.19) *n* = 62			10.74 (3.55) *n* = 76		
Control						Experimental					
10.97 (4.01) *n* = 99						11.11 (3.86) *n* = 138					
On-Level		Gifted		Sp. Needs		On-Level		Gifted		Sp. Needs	
11.01 (3.77) *n* = 71		9.11 (3.80) *n* = 18		14.00 (4.50) *n* = 10		11.23 (3.68) *n* = 113		10.25 (4.68) *n* = 16		11.11 (4.68) *n* = 9	
A	B	A	B	A	B	A	B	A	B	A	B
11.09 (3.45) *n* = 23	10.98 (3.95) *n* = 48	9.11 (3.80) *n* = 18	-- -- *n* = 0	15.75 (2.22) *n* = 4	12.83 (5.42) *n* = 6	12.29 (3.76) *n* = 38	10.69 (3.55) *n* = 75	10.25 (4.68) *n* = 16	-- -- *n* = 0	10.75 (4.86) *n* = 8	14.00 -- *n* = 1

Initial Desire to Do Science Scores

Teacher A			Teacher B			Teacher A			Teacher B		
22.13 (5.07) *n* = 45			21.15 (5.52) *n* = 54			19.95 (5.76) *n* = 62			21.07 (5.56) *n* = 76		
Control						Experimental					
21.60 (5.31) *n* = 99						20.57 (5.66) *n* = 138					
On-Level		Gifted		Sp. Needs		On-Level		Gifted		Sp. Needs	
21.63 (5.37) *n* = 71		22.67 (5.10) *n* = 18		19.40 (5.13) *n* = 10		20.51 (5.57) *n* = 113		20.31 (6.29) *n* = 16		21.67 (6.14) *n* = 9	
A	B	A	B	A	B	A	B	A	B	A	B
22.39 (4.85) *n* = 23	21.27 (5.62) *n* = 48	22.67 (5.10) *n* = 18	-- -- *n* = 0	18.25 (5.91) *n* = 4	20.17 (4.96) *n* = 6	19.50 (5.47) *n* = 38	21.03 (5.59) *n* = 75	20.31 (6.29) *n* = 16	-- -- *n* = 0	21.38 (6.50) *n* = 8	24.00 -- *n* = 1

Initial Perception of the Science Teacher Scores

Teacher A			Teacher B			Teacher A			Teacher B		
11.93 (2.09) *n* = 45			11.44 (2.38) *n* = 54			11.65 (2.23) *n* = 62			11.41 (2.09) *n* = 76		
Control						Experimental					
11.67 (2.25) *n* = 99						11.51 (2.15) *n* = 138					
On-Level		Gifted		Sp. Needs		On-Level		Gifted		Sp. Needs	
11.72 (2.46) *n* = 71		11.89 (1.60) *n* = 18		10.90 (1.66) *n* = 10		11.51 (2.02) *n* = 113		11.00 (2.76) *n* = 16		12.44 (2.51) *n* = 9	
A	B	A	B	A	B	A	B	A	B	A	B
12.09 (2.54) *n* = 23	11.54 (2.42) *n* = 48	11.89 (1.60) *n* = 18	-- -- *n* = 0	11.25 (1.26) *n* = 4	10.67 (1.97) *n* = 6	11.74 (1.86) *n* = 38	11.40 (2.11) *n* = 75	11.00 (2.76) *n* = 16	-- -- *n* = 0	12.50 (2.67) *n* = 8	12.00 -- *n* = 1

Hypothesis #1: Explicit Understanding

Previously, this study hypothesized that engagement with the tutoring system would improve students' declarative knowledge of the skills involved in inquiry-driven modeling, including an understanding of the scientific method, an understanding of the process of modeling and inquiry, an understanding of model evaluation criteria, and an understanding the importance of examining invisible components in modeling ecological systems. Toward this end, students completed a content test on these topics before and after the experiment. This content test touched on all of these elements in fifteen multiple-choice questions administered concurrently with the attitudinal inventories and can be seen in Appendix B.

Initial t-test analysis revealed that there was an overall improvement in students' performance on the content test through the intervention (t = 3.48, *p* < 0.001). Students prior to the intervention scored an average of 9.46, while students after the intervention scored an average of 10.44. The 95% confidence interval surrounding the difference in performance between the pre- and post-test is (0.430, 1.545), indicating an improvement of between half of a question and one and a half questions. Following this analysis, a multivariate ANOVA was conducted to determine the specific impact of each of the three independent variables on the changes in students' content knowledge through the intervention. As in the preliminary analysis described above, multivariate ANOVA was conducted simultaneously on content scores and all five attitudinal metrics in order to reduce the risk of Type I error. Table 25 displays the multivariate ANOVA analysis for all six of these variables (I.2), followed by the ANOVA analysis specifically for the change in students' content scores (I.2.A).

Table 25: Multivariate ANOVA results on the change in students' content and attitudinal scores over the course of the intervention. The initial multivariate ANOVA points to an effect of the experimental condition on the change in performance, but the subsequent ANOVA on the change in content scores alone reveals that the effect of the experimental condition is limited to the attitudinal metrics (discussed in the next section). A relationship is observed, however, between the Teacher and the change in content scores, with students in Teacher B's class improving their scores significantly more than in Teacher A's class.

```
Multivariate ANOVA :

                         Df   Pillai approx F num Df den Df  Pr(>F)
Condition                 1 0.063999  2.52985      6    222 0.02176
Teacher                   1 0.060708  2.39137      6    222 0.02933
Level                     2 0.043402  0.82444     12    446 0.62523
Condition:Teacher         1 0.009625  0.35959      6    222 0.90378
Condition:Level           2 0.046142  0.87772     12    446 0.56988
Teacher:Level             1 0.031328  1.19664      6    222 0.30904
Condition:Teacher:Level   1 0.019681  0.74283      6    222 0.61570
Residuals               227

Response Content.D :
                         Df  Sum Sq Mean Sq F value   Pr(>F)
Condition                 1    2.20   2.197  0.3886 0.533638
Teacher                   1   43.81  43.806  7.7500 0.005824
Level                     2    7.68   3.839  0.6791 0.508092
Condition:Teacher         1    1.54   1.539  0.2722 0.602353
Condition:Level           2    4.29   2.143  0.3792 0.684845
Teacher:Level             1    0.47   0.471  0.0833 0.773140
Condition:Teacher:Level   1    1.90   1.903  0.3367 0.562323
Residuals               227 1283.08   5.652
```

The initial multivariate ANOVA reveals that there exist significant relationships between both the condition and the teacher on the change in students' scores on the content tests or attitudinal metrics. The subsequent ANOVA specifically on the change in students' content scores reveals that only the teacher has a statistically significant influence on the change in students' content scores (F = 7.75, p < 0.01). Students in

Teacher B's class improve significantly more in their performance on the content test compared to students in Teacher A's class. It is worth noting that this is the opposite of the relationship observed in the preliminary analysis covered earlier: on the pre-test, the students in Teacher A's class outperformed the students in Teacher B's class with statistical significance (F = 2.32, $p < 0.01$). Here, students in Teacher B's class improved more than students in Teacher A's class with statistical significance (F = 7.75, $p < 0.01$), leading to no statistically significant difference between students in Teacher B's class and students in Teacher A's class in performance on the post-test. A relationship still exists, however, between Level and performance on the post-test.

Table 26: Difference between post-test and pre-test content scores for students in each group and subgroup. The top number in each cell is the mean for that group; the number in parentheses is the standard deviation within that group; the bottom number is the number of samples within that subgroup. See Table 22 for information about the groups of students summarized in each cell.

Difference between Post-Test and Pre-Test Content Scores

<table>
<tr><td colspan="3">Teacher A</td><td colspan="3">Teacher B</td><td colspan="3">Teacher A</td><td colspan="3">Teacher B</td></tr>
<tr><td colspan="3">+0.53
(2.26)
n = 45</td><td colspan="3">+1.57
(2.17)
n = 54</td><td colspan="3">+0.50
(2.35)
n = 62</td><td colspan="3">+1.24
(2.55)
n = 76</td></tr>
<tr><td colspan="6">Control</td><td colspan="6">Experimental</td></tr>
<tr><td colspan="6">+1.10
(2.26)
n = 99</td><td colspan="6">+0.91
(2.48)
n = 138</td></tr>
<tr><td colspan="2">On-Level</td><td colspan="2">Gifted</td><td colspan="2">Sp. Needs</td><td colspan="2">On-Level</td><td colspan="2">Gifted</td><td colspan="2">Sp. Needs</td></tr>
<tr><td colspan="2">+1.27
(2.46)
n = 71</td><td colspan="2">+0.56
(1.50)
n = 18</td><td colspan="2">+0.90
(1.91)
n = 10</td><td colspan="2">+1.11
(2.48)
n = 113</td><td colspan="2">+0.06
(2.14)
n = 16</td><td colspan="2">−0.11
(2.76)
n = 9</td></tr>
<tr><td>A</td><td>B</td><td>A</td><td>B</td><td>A</td><td>B</td><td>A</td><td>B</td><td>A</td><td>B</td><td>A</td><td>B</td></tr>
<tr><td>+0.48
(2.89)
n = 23</td><td>+1.65
(2.15)
n = 48</td><td>+0.56
(1.50)
n = 18</td><td>--
--
n = 0</td><td>+0.75
(0.96)
n = 4</td><td>+1.00
(2.45)
n = 6</td><td>+0.84
(2.31)
n = 38</td><td>+1.24
(2.56)
n = 75</td><td>+0.06
(2.14)
n = 16</td><td>--
--
n = 0</td><td>−0.25
(2.92)
n = 8</td><td>+1.00
--
n = 1</td></tr>
</table>

In summary, based on overall results from the pre- and post-test covering declarative knowledge of the inquiry-driven modeling process**, no statistically significant effect was observed on students' performance on the content test from**

exposure to the tutoring system. There did exist statistically significant improvement in performance overall, as well as statistically significant improvement based on the teacher, but no effect of the tutoring system was observed. Thus, **Hypothesis #1 is not accepted**. There are **not** sufficient data to conclude that access to MILA–T leads to improved content knowledge compared to participation in the intervention without MILA–T.

Hypothesis #2: Dispositional Framing

Students' attitudes were assessed according to five different established attitudinal metrics before and after the intervention. Two of these, students' attitudes toward scientific inquiry and interest in science as a career, come from the Test of Science Related Attitudes (Fraser 1981). Fraser describes Attitude toward Scientific Inquiry as "acceptance of scientific inquiry as a way of thought", and Career Interest in Science as "development of interest in pursuing a career in science". Each construct is determined by responses to ten different items, evenly split between positive and negative presentations of the underlying attitude. For example, a negative item for Career Interest in Science is, "A career in science would be dull and boring", while a positive item for Attitude toward Scientific Inquiry is, "I would prefer to find out why something happens by doing an experiment than by being told."

The other three come from the Modified Attitudes Towards Science Inventory (Weinburgh & Steele 2000), and are students' anxiety toward science, perception of the science teacher, and desire to do science. The first, Anxiety toward Science, captures the degree to which students regard themselves as unfit or uninterested in science; for example, one item offers, "I feel tense when someone talks to me about science". The second, Perception of the Science Teacher, looks at what role students see the teacher playing in the classroom; for example, "Science teachers present material in a clear way". The third, Desire to Do Science, examines the extent to which students are interested in engaging in scientific activities outside their required assignments; for example, "Sometimes I read ahead in our science book".

As referenced previously, no statistically significant differences were observed in students according to these attitudinal metrics prior to the intervention based on assignment to the Control or Experimental condition. Statistically significant differences were observed, however, due the students' Levels. Gifted students expressed higher career interest in science and lower anxiety toward science than On-Level students, who in turn performed more positively than Special Needs students according to both metrics as well. Special Needs students, however, reported a statistically significantly higher attitude toward science inquiry than Gifted students, who in turn reported a higher response than On-Level students.

An initial multivariate ANOVA, shown earlier in Table 25 and displayed again below in Table 27 for reference, showed a statistically significant relationship between both Condition and Teacher and some parts of the change in either content or attitudinal scores (I.2). Subsequent ANVOA analysis, described in the remainder of this section, shows a statistically significant effect of Condition on Attitude toward Scientific Inquiry (I.2.B), as well as a borderline significant effect of Condition on Career Interest in Science (I.2.C). Additional relationships between Teacher and other attitudinal scores are also present.

Table 27: Multivariate ANOVA results on the change in students' content and attitudinal scores over the course of the intervention. The initial multivariate ANOVA points to effects of the experimental condition and teacher on the change in performance (*p* < 0.05).

```
                         Df   Pillai approx F num Df den Df  Pr(>F)
Condition                 1 0.063999  2.52985      6    222 0.02176
Teacher                   1 0.060708  2.39137      6    222 0.02933
Level                     2 0.043402  0.82444     12    446 0.62523
Condition:Teacher         1 0.009625  0.35959      6    222 0.90378
Condition:Level           2 0.046142  0.87772     12    446 0.56988
Teacher:Level             1 0.031328  1.19664      6    222 0.30904
Condition:Teacher:Level   1 0.019681  0.74283      6    222 0.61570
Residuals               227
```

Change in Attitude toward Scientific Inquiry

As established previously, prior to the intervention, no statistically significant difference existed between Attitudes toward Scientific Inquiry in the Control and Experimental groups. After the multivariate ANOVA indicated a statistically significant effect of Condition on change in some constructs, an ANOVA was run specifically on this change to identify what relationship, if any, existed with regard to Attitude toward Scientific Inquiry. This ANOVA, shown below in Table 28, showed a statistically significant effect of Condition on change in Attitude toward Scientific Inquiry (F = 8.59, p < 0.01) (I.2.B).

Table 28: Univariate ANOVA results on the change in students' Attitudes toward Scientific Inquiry over the course of the intervention. These results show a statistically significant effect of Condition on change in Attitude toward Scientific Inquiry between the pre-test and post-test.

```
 Response AtSI.D :
                         Df  Sum Sq Mean Sq F value   Pr(>F)
Condition                 1   397.3  397.33  8.5903 0.003725
Teacher                   1   118.7  118.74  2.5672 0.110490
Level                     2    58.6   29.29  0.6333 0.531748
Condition:Teacher         1     0.7    0.74  0.0160 0.899406
Condition:Level           2   210.0  104.99  2.2698 0.105676
Teacher:Level             1    52.4   52.37  1.1323 0.288419
Condition:Teacher:Level   1    15.7   15.75  0.3405 0.560140
Residuals               227 10499.6   46.25
```

Full results on the change in students' Attitude toward Scientific Inquiry for each group are shown below in Table 29. Students in the Experimental group experienced an average increase of 1.46 points on their attitude toward scientific inquiry score (σ = 7.16). Students in the Control group, on the other hand, experienced an average *decrease* of 1.16 points on their attitude toward scientific inquiry score (σ = 6.33). Thus, students who had access to MILA–T experienced a statistically significant increase in Attitude toward Scientific Inquiry relative to students who did not have access to MILA–T.

Table 29: Difference between post-test and pre-test Attitude toward Scientific Inquiry scores for students in each group and subgroup. The top number in each cell is the mean for that group; the number in parentheses is the standard deviation within that group; the bottom number is the number of samples within that subgroup. See Table 22 for further information about the groups of students summarized in each cell.

Difference between Post-Test and Pre-Test Attitude toward Scientific Inquiry Scores

Teacher A			Teacher B			Teacher A			Teacher B		
−2.07 (6.50) *n* = 45			−0.41 (6.15) *n* = 54			+0.77 (8.27) *n* = 62			+2.03 (6.11) *n* = 76		
Control						Experimental					
−1.16 (6.33) *n* = 99						+1.46 (7.16) *n* = 138					
On-Level		Gifted		Sp. Needs		On-Level		Gifted		Sp. Needs	
−1.27 (6.81) *n* = 71		−1.89 (5.41) *n* = 18		+0.90 (3.78) *n* = 10		+2.12 (6.61) *n* = 113		−1.62 (10.50) *n* = 16		−1.33 (5.17) *n* = 9	
A	B	A	B	A	B	A	B	A	B	A	B
−3.04 (7.52) *n* = 23	−0.42 (6.36) *n* = 48	−1.89 (5.41) *n* = 18	-- -- *n* = 0	+2.75 (0.50) *n* = 4	−0.33 (4.59) *n* = 6	+2.21 (7.53) *n* = 38	+2.08 (6.14) *n* = 75	−1.62 (10.50) *n* = 16	-- -- *n* = 0	−1.25 (5.52) *n* = 8	−2.00 -- *n* = 1

Given the statistically significant relationship between Condition and change in Attitude toward Scientific Inquiry, a subsequent t-test analysis was conducted on the data to identify the exact nature of the change in Attitude toward Scientific Inquiry. T-test analysis of the difference between pre-test and post-test Attitude toward Scientific Inquiry scores for the Control group revealed no statistically significant difference (t = 1.16, p = 0.27); thus, while on average Attitude toward Scientific Inquiry dropped in the Control group, this drop was not statistically significant on its own. T-test analysis of the difference between pre-test and post-test Attitude toward Scientific Inquiry for the Experimental group revealed a borderline statistically significant difference (t = 1.65, $p <$ 0.10). Finally, t-test analysis on the difference between Attitude toward Scientific Inquiry in Control and Experimental students (that is, not the change in scores, but the difference in the groups' final scores) shows **students in the Experimental condition closed the study with a significantly higher Attitude toward Scientific Inquiry than students in**

the Control condition (t = 2.12, *p* < 0.05). Prior to the study, there existed no statistically significant difference in Attitude toward Scientific Inquiry between Control and Experimental students; thus, this further confirms that access to MILA–T caused a positive and statistically significant effect on Attitude toward Scientific Inquiry compared to participation in the project without MILA–T. It should be noted that this measurement was taken immediately after the study closed, however; it is possible, and in fact probable, that this difference between Control and Experimental groups at the close of the study would fade over time. The takeaway of this analysis is that interaction with a metacognitive tutoring system during this activity improves students' attitudes in the short term.

Change in Career Interest in Science

As established previously, prior to the intervention, no statistically significant difference existed between Career Interest in Science in the Control and Experimental groups. After the multivariate ANOVA indicated a statistically significant effect of Condition on change in some constructs, an ANOVA was run specifically on this change to identify what relationship, if any, existed with regard to Career Interest in Science. This ANOVA, shown below in Table 30, showed a borderline statistically significant effect of Condition on change in Career Interest in Science (F = 3.22, *p* < 0.1) (I.2.C).

Table 30: Univariate ANOVA results on the change in students' Career Interest in Science over the course of the intervention. These results show a borderline significant effect of Condition on change in Career Interest in Science between the pre-test and post-test.

```
Response CIiS.D :
                         Df Sum Sq Mean Sq F value Pr(>F)
Condition                 1  130.3 130.349  3.2238 0.0739
Teacher                   1    0.0   0.003  0.0001 0.9929
Level                     2   67.2  33.588  0.8307 0.4371
Condition:Teacher         1    6.0   5.973  0.1477 0.7011
Condition:Level           2   19.3   9.652  0.2387 0.7878
Teacher:Level             1  192.6 192.589  4.7632 0.0301
Condition:Teacher:Level   1    7.3   7.297  0.1805 0.6714
Residuals               227 9178.2  40.433
```

Full results on the change in students' Attitude toward Scientific Inquiry for each group are shown below in Table 31. Students in the Experimental group experienced an average increase of 2.03 points on their Career Interest in Science score (σ = 6.01), while students in the Control group experienced an average increase of 0.53 points on their Career Interest in Science score (σ = 6.79). Given the positive change in scores in both the Control and Experimental groups, a follow-up t-test was conducted to establish the significance of the change. This t-test found no statistically significant difference between pre-test and post-test scores for students in the Control group (t = 0.44, *p* = 0.66), but it did demonstrate a statistically significant difference between pre-test and post-test scores for students in the Experimental group (t = 1.98, *p* < 0.05) (I.3).

Table 31: Difference between post-test and pre-test Career Interest in Science scores for students in each group and subgroup. The top number in each cell is the mean for that group; the number in parentheses is the standard deviation within that group; the bottom number is the number of samples within that subgroup. See Table 22 for further information about the groups of students summarized in each cell.

Difference between Post-Test and Pre-Test Career Interest in Science Scores

Teacher A			Teacher B			Teacher A			Teacher B		
+0.67 (5.12) *n* = 45			+0.41 (7.96) *n* = 54			+1.92 (6.20) *n* = 62			+2.12 (5.89) *n* = 76		
Control						Experimental					
+0.53 (6.79) *n* = 99						+2.03 (6.01) *n* = 138					
On-Level		Gifted		Sp. Needs		On-Level		Gifted		Sp. Needs	
+0.15 (7.01) *n* = 71		+0.56 (5.63) *n* = 18		+3.10 (7.17) *n* = 10		+1.92 (6.18) *n* = 113		+2.44 (5.81) *n* = 16		+2.67 (4.53) *n* = 9	
A	B	A	B	A	B	A	B	A	B	A	B
+1.13 (4.97) *n* = 23	−0.31 (7.80) *n* = 48	+0.56 (5.63) *n* = 18	-- -- *n* = 0	−1.50 (4.04) *n* = 4	+6.17 (7.39) *n* = 6	+1.68 (6.79) *n* = 38	+2.04 (5.89) *n* = 75	+2.44 (5.81) *n* = 16	-- -- *n* = 0	+2.00 (4.34) *n* = 8	+8.00 -- *n* = 1

This analysis shows that students in the Experimental group experienced a significant improvement in Career Interest in Science that students in the Control group did not experience. **Students in the Experimental group showed a statistically significant improvement in Career Interest in Science, while students in the Control group did not**. The weak effect identified by the initial ANOVA renders this conclusion less conclusive than the observed change in Attitude toward Scientific Inquiry, but a change was observed nonetheless. The previous note regarding the permanence of these attitudinal changes applies here as well; this observation was made immediately following the close of the study, and the observed effect is likely to fade over time. The new challenge would be to identify the degree to which the change is permanent, and to structure new interventions using this system to help make the change more permanent.

Change in Anxiety toward Science

On the Anxiety toward Science metric, higher scores correspond to more anxiety toward science; therefore, a drop in Anxiety toward Science score is considered a desirable result. As established previously, prior to the intervention, no statistically significant difference existed between Anxiety toward Science in the Control and Experimental groups. After the multivariate ANOVA indicated a statistically significant effect of Condition on change in some constructs, an ANOVA was run specifically on this change to identify what relationship, if any, existed with regard to Anxiety toward Science. This ANOVA, shown below in Table 32, shows that there is **no statistically significant relationship between change in Anxiety toward Science and participation in the Experimental condition**. A relationship does exist, however, between the Teacher and the change in Anxiety toward Science (F = 5.77, $p < 0.05$) (I.2.D).

Further analysis revealed that although each teachers' students dropped in their Anxiety toward Science, Teacher A's students dropped more (2.52 points, σ = 3.17) than Teacher B's students (1.48 points, σ = 3.41). Table 33 below shows scores across all groups and subgroups; no significant effects are observed due to any factor except Teacher.

Table 32: Univariate ANOVA results on the change in students' Attitudes toward Scientific Inquiry over the course of the intervention. These results show a statistically significant effect of Condition on change in Attitude toward Scientific Inquiry between the pre-test and post-test.

```
Response AtS.D :
                         Df  Sum Sq Mean Sq F value  Pr(>F)
Condition                 1    0.00   0.000  0.0000 0.99960
Teacher                   1   64.27  64.272  5.7720 0.01709
Level                     2    3.49   1.747  0.1569 0.85491
Condition:Teacher         1    9.28   9.284  0.8337 0.36217
Condition:Level           2   13.72   6.858  0.6159 0.54104
Teacher:Level             1    0.01   0.012  0.0011 0.97373
Condition:Teacher:Level   1   10.97  10.968  0.9850 0.32203
Residuals               227 2527.65  11.135
```

In addition to students in Teacher A's classes improving more in their Anxiety toward Science than students in Teacher B's classes, a significant improvement was observed overall. On the pre-survey, students replied with an average score of 11.05 (σ = 3.91). On the post-survey, students replied with an average score of 9.10 (σ = 2.94). This change was identical across the Control and Experimental conditions, and although the drop was larger in Teacher A's classes than in Teacher B's, the drop was statistically significant within each class.

Given that this change occurred identically across the Control and Experimental groups, it is reasonable to speculate that the cause of this change was not within the intervention itself. One likely explanation is that between the beginning and end of the intervention, students received their scores on the standardized End of Course Tests. Informal conversation suggested that students across all classes were pleased with their scores, and that students in Teacher A's classes generally performed better than students in Teacher B's. Given this positive performance on a crucial standardized test, it is reasonable to infer that this development – or perhaps another event that affected all classes – could be responsible for this significant change. Thus, although the drop in

Anxiety toward Science across both Teachers and Conditions is notable, it cannot be attributed to this intervention.

Table 33: Difference between post-test and pre-test Anxiety toward Science scores for students in each group and subgroup. The top number in each cell is the mean for that group; the number in parentheses is the standard deviation within that group; the bottom number is the number of samples within that subgroup. See Table 22 for further information about the groups of students summarized in each cell.

Difference between Post-Test and Pre-Test Anxiety toward Science Scores

Teacher A			Teacher B			Teacher A			Teacher B		
−2.78 (2.54) *n* = 45			−1.26 (3.20) *n* = 54			−2.34 (3.57) *n* = 62			−1.63 (3.56) *n* = 76		
Control						Experimental					
−1.95 (3.00) *n* = 99						−1.95 (3.57) *n* = 138					
On-Level		Gifted		Sp. Needs		On-Level		Gifted		Sp. Needs	
−1.73 (2.89) *n* = 71		−2.56 (2.23) *n* = 18		−2.40 (4.74) *n* = 10		−1.97 (3.43) *n* = 113		−2.19 (4.31) *n* = 16		−1.22 (4.21) *n* = 9	
A	B	A	B	A	B	A	B	A	B	A	B
−2.74 (2.63) *n* = 23	−1.25 (2.91) *n* = 48	−2.56 (2.23) *n* = 18	-- -- *n* = 0	−4.00 (3.65) *n* = 4	−1.33 (5.39) *n* = 6	−2.68 (3.03) *n* = 38	−1.61 (3.58) *n* = 75	−2.19 (4.31) *n* = 16	-- -- *n* = 0	−1.00 (4.44) *n* = 8	−3.00 -- *n* = 1

Change in Desire to Do Science

As established previously, prior to the intervention, no statistically significant difference existed between Desire to Do Science in the Control and Experimental groups. After the multivariate ANOVA indicated a statistically significant effect of Condition on change in some constructs, an ANOVA was run specifically on this change to identify what relationship, if any, existed with regard to Desire to Do Science. This ANOVA, shown below in Table 34, showed that **there is no relationship between any independent variable and change in Desire to Do Science**.

Table 34: Univariate ANOVA results on the change in students' Attitudes toward Scientific Inquiry over the course of the intervention. These results show a statistically significant effect of Condition on change in Attitude toward Scientific Inquiry between the pre-test and post-test.

```
Response DtDS.D :
                          Df Sum Sq Mean Sq F value Pr(>F)
Condition                  1    6.5   6.532  0.2912 0.5900
Teacher                    1   27.8  27.812  1.2399 0.2667
Level                      2    0.2   0.121  0.0054 0.9946
Condition:Teacher          1    5.4   5.440  0.2425 0.6229
Condition:Level            2    6.9   3.470  0.1547 0.8568
Teacher:Level              1   41.1  41.146  1.8343 0.1770
Condition:Teacher:Level    1    0.5   0.452  0.0202 0.8872
Residuals                227 5091.9  22.431
```

The full results across all conditions are displayed below in Table 35. Although minor effects seem present due to Condition, Teacher, and Level, none of these differences are statistically significant. Thus, there are insufficient data to conclude that any factors held a relationship with change in Desire to Do Science (I.2.E).

Table 35: Difference between post-test and pre-test Desire to Do Science scores for students in each group and subgroup. The top number in each cell is the mean for that group; the number in parentheses is the standard deviation within that group; the bottom number is the number of samples within that subgroup. See Table 22 for further information about the groups of students summarized in each cell.

Difference between Post-Test and Pre-Test Desire to Do Science Scores

Teacher A			Teacher B			Teacher A			Teacher B		
+0.27 (3.76) *n* = 45			−0.06 (4.35) *n* = 54			+0.95 (5.64) *n* = 62			+0.00 (4.58) *n* = 76		
Control						Experimental					
+0.09 (4.07) *n* = 99						+0.43 (5.09) *n* = 138					
On-Level		Gifted		Sp. Needs		On-Level		Gifted		Sp. Needs	
+0.14 (4.25) *n* = 71		−0.06 (2.48) *n* = 18		+0.00 (5.35) *n* = 10		+0.26 (4.71) *n* = 113		+1.31 (7.88) *n* = 16		+1.00 (3.67) *n* = 9	
A	B	A	B	A	B	A	B	A	B	A	B
+0.87 (4.50) *n* = 23	−0.21 (4.13) *n* = 48	−0.06 (2.48) *n* = 18	-- -- *n* = 0	−1.75 (3.86) *n* = 4	+1.17 (6.21) *n* = 6	+0.84 (4.95) *n* = 38	−0.04 (4.60) *n* = 75	+1.31 (7.88) *n* = 16	-- -- *n* = 0	+0.75 (3.85) *n* = 8	+3.00 -- *n* = 1

In addition to the absence of data suggesting a relationship between any independent variable and change in Desire to Do Science, no data exist to suggest a change in Desire to Do Science overall. Students opened the study with an average Desire to Do Science score of 20.99 (σ = 5.53) and closed the student with an average Desire to Do Science score of 21.28 (σ = 5.76), a statistically insignificant (t = 0.56, *p* = 0.58) change of +0.27. Thus, no data exist to suggest that any facet of the study had any impact on the Desire to Do Science metric.

Change in Perception of the Science Teacher

As established previously, prior to the intervention, no statistically significant difference existed between Perception of the Science Teacher in the Control and Experimental groups. After the multivariate ANOVA indicated a statistically significant effect of Condition on change in some constructs, an ANOVA was run specifically on this change to identify what relationship, if any, existed with regard to Perception of the Science Teacher. This ANOVA, shown below in Table 36, showed that **there is no statistically significant relationship between any independent variable and change in Perception of the Science Teacher** (I.2.F).

Table 36: Univariate ANOVA results on the change in students' Attitudes toward Scientific Inquiry over the course of the intervention. These results show a statistically significant effect of Condition on change in Attitude toward Scientific Inquiry between the pre-test and post-test.

```
Response PofTS.D :
                         Df  Sum Sq Mean Sq F value  Pr(>F)
Condition                 1    3.40  3.4001  0.6073 0.43661
Teacher                   1   20.39 20.3895  3.6420 0.05760
Level                     2   13.22  6.6101  1.1807 0.30894
Condition:Teacher         1    0.28  0.2754  0.0492 0.82468
Condition:Level           2    3.90  1.9503  0.3484 0.70622
Teacher:Level             1   23.19 23.1908  4.1424 0.04299
Condition:Teacher:Level   1   11.44 11.4417  2.0437 0.15421
Residuals               227 1270.85  5.5985
```

A borderline statistically significant relationship does exist between Teacher and change in Perception of the Science Teacher. Students in Teacher A's classes experienced no change to their Perception of the Science Teacher scores (dropping 0.01 from 11.77 to

11.76), students in Teacher B's classes reported a drop of 0.60 in their Perception of the Science Teacher, from 11.42 to 10.82. On its own, this drop was statistically significant. An interaction was also indicated between Teacher and Level, but the initial multivariate ANOVA did not indicate any significant effect of this interaction, suggesting that this interaction is merely a product of repeated testing.

Table 37: Difference between post-test and pre-test Perception of the Science Teacher scores for students in each group and subgroup. The top number in each cell is the mean for that group; the number in parentheses is the standard deviation within that group; the bottom number is the number of samples within that subgroup. See Table 22 for further information about the groups of students summarized in each cell.

Difference between Post-Test and Pre-Test Perception of the Science Teacher Scores

Teacher A			Teacher B			Teacher A			Teacher B		
+0.20 (2.39) *n* = 45			−0.52 (2.55) *n* = 54			−0.16 (2.44) *n* = 62			−0.66 (2.21) *n* = 76		
Control						Experimental					
−0.19 (2.49) *n* = 99						−0.43 (2.32) *n* = 138					
On-Level		Gifted		Sp. Needs		On-Level		Gifted		Sp. Needs	
−0.14 (2.67) *n* = 71		−0.11 (1.41) *n* = 18		−0.70 (2.83) *n* = 10		−0.43 (2.09) *n* = 113		+0.00 (3.52) *n* = 16		−1.22 (2.54) *n* = 9	
A	B	A	B	A	B	A	B	A	B	A	B
+0.87 (2.82) *n* = 23	−0.62 (2.48) *n* = 48	−0.11 (1.41) *n* = 18	-- -- *n* = 0	−2.25 (1.50) *n* = 4	+0.33 (3.14) *n* = 6	−0.03 (1.78) *n* = 38	−0.64 (2.22) *n* = 75	+0.00 (3.52) *n* = 16	-- -- *n* = 0	−1.12 (2.70) *n* = 8	−2.00 -- *n* = 1

Taken overall, no significant effect on Perception of the Science Teacher was observed. Students reported a score of 11.58 on the pre-survey (σ = 2.19) and 11.24 on the post-survey (σ = 2.44), a statistically insignificant change of −0.34 (t = 1.55, p = 0.12).

Summary of Change in Dispositions

The change in content test scores and attitudinal scores across the five metrics are summarized below in Figure 23.

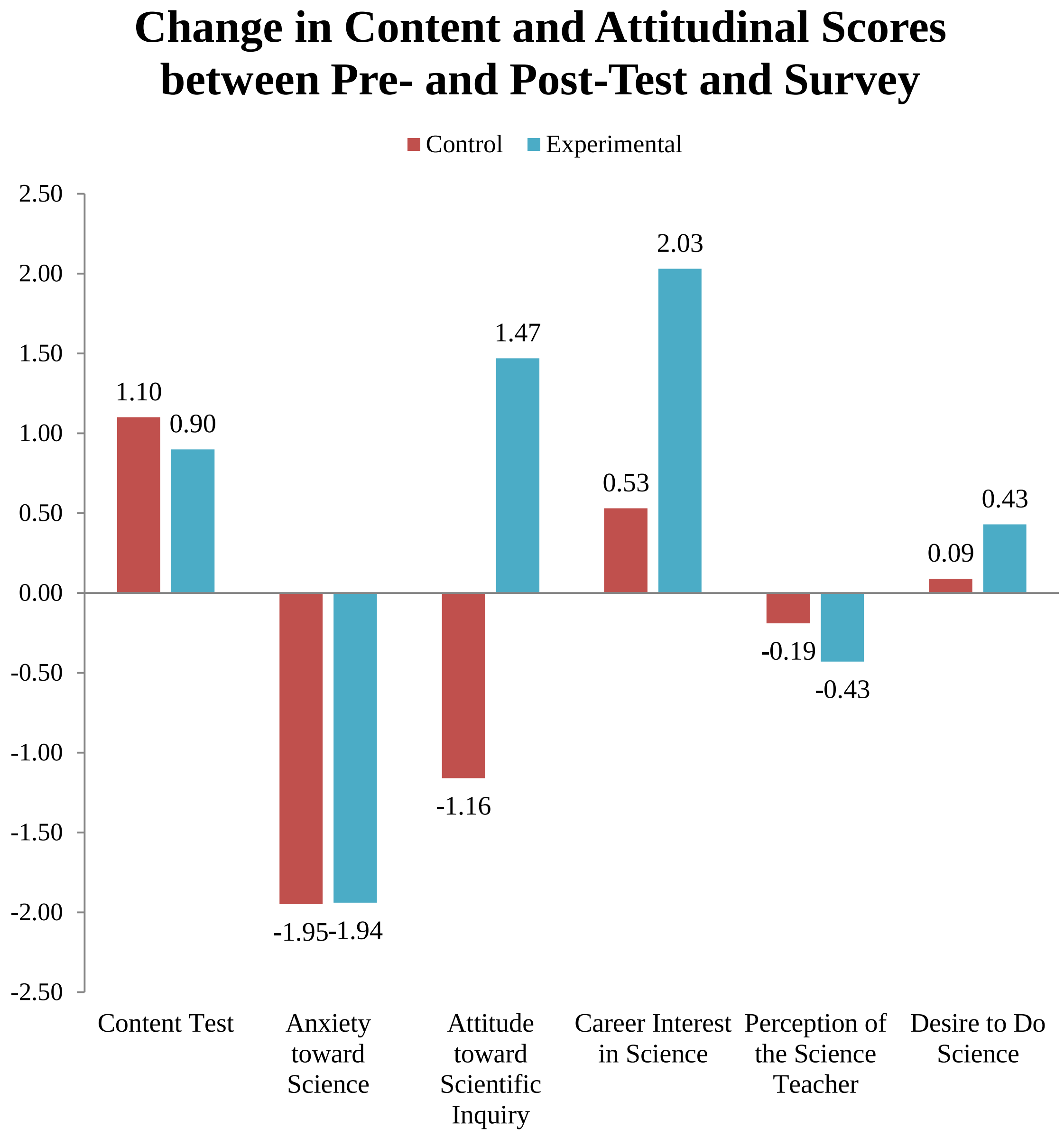


Figure 23: The average changes in content test scores and attitudinal survey scores in the Control and Experimental groups. Attitude toward Scientific Inquiry and Career Interest in Science exhibited a statistically significant interaction between the condition and the change in scores. Content Test and Anxiety toward Science exhibited statistically significant overall changes, but no statistically significant interaction with condition. Perception of the Science Teacher and Desire to Do Science exhibited neither overall change nor interaction with the condition.

The results presented in this section establish three important changes as a result of the intervention. First, a statistically significant relationship (F = 8.59, $p < 0.01$) exists between participation in the experimental condition and the change in students' Attitudes toward Scientific Inquiry: students in the Experimental condition experienced a positive and statistically significant change in Attitude toward Scientific Inquiry compared to students in the Control condition, leading to a positive and statistically significant difference in final Attitude toward Scientific Inquiry for students in the Experimental condition (t = 2.12, $p < 0.05$). Second, students in the Experimental group experienced a statistically significant increase in Career Interest in Science (t = 1.98, $p < 0.05$), while students in the Control group experienced no such statistically significant increase (t = 0.44, p > 0.1); this arose from a borderline significant relationship (F = 3.22, $p < 0.1$) between participation in the experimental condition and change in Career Interest in Science. Third, students in both the Control and the Experimental conditions experienced statistically significant decreases in Anxiety toward Science, although no difference between the groups was observed.

The motivating hypothesis behind this analysis was that access to MILA–T during participation in the intervention would lead to a positive effect on attitudes toward science. The first result above, the change in Attitude toward Scientific Inquiry, confirms this hypothesis; students with MILA–T experienced a significant improvement in Attitude toward Scientific Inquiry compared to students that did not have access to MILA–T. The second result above, the change in Career Interest in Science, further supports this hypothesis, although the borderline significant relationship between participation in the experimental condition and change in Career Interest in Science leaves this result slightly less conclusive on its own. Based on these two results together, **Hypothesis #2 is accepted**. There **are** sufficient data to conclude that access to MILA–T leads to improved attitudes toward science compared to participation in the intervention without MILA–T. These results can also be found in Joyner & Goel 2014.

Additional Results

The analyses presented in the previous sections address students' understanding of and dispositions toward inquiry-driven modeling. In addition to the data used in these analyses, additional data were collected on other components and developments in the classroom. Although these do not explicitly address the original hypotheses, these help provide a more complete view of the intervention, as well as contextualize some of the specific developments. Here, two such topics are discussed: changes to students' class interests across the intervention and students' perceptions of both MILA itself and MILA–T.

Changes to Students' Class Interests After Intervention

As part of the attitudinal survey administered at the beginning and end of the intervention, students were asked to mark what classes they were interested in taking in high school. Students were given twelve options: astronomy, earth science, geometry, trigonometry, physics, environmental science, biology, geobiology, calculus, chemistry, algebra, and astrotrigology. The final subject is fictitious, but intended to capture the possibility of general interest in science-related classes regardless of personal familiarity. For each course, each student was assessed a score of −1, 0, or 1; a −1 represents a course that the student stated they were interested in at the start, but not at the finish. A 1 represents a course that the student was not interested in at the start, but was interested in at the end. A 0 represents a course that was unchanged; either it was not marked in either the pre- or post-survey, or it was marked in both. These scores were then added in four groups: "All", the overall change in the student's course interest (all twelve scores); "Science", the change in the student's interest in science courses (astronomy, earth science, physics, environmental science, biology, geobiology, chemistry); "Math", the change in the student's interest in math courses (geometry, trigonometry, calculus, algebra); and "Life Science", the change in the student's interest in specifically life science courses (biology, geobiology, environmental science).

First, a straightforward repeated t-test revealed no evidence that students' overall preferences changed. In the "All" category of classes, students experience an average change of −0.004 classes (σ = 1.784), representing effectively no change in class interest. The "Science", "Math", and "Life Science" categories saw changes of −0.025 (σ = 1.378), −0.021 (σ = 0.936), and −0.013 (σ = 0.628) respectively, showing no effective change in those three categories as well. A multivariate analysis of variance was conducted to check for any interactions between the independent variables (Condition, Teacher, and Level) and the individual students' change in class interests. The results of this multivariate analysis of variance are below.

Table 38: Multivariate ANOVA results on the interaction between Condition, Teacher, and Level and students' change in class interest over the course of the intervention. No significant interactions were observed.

```
                         Df   Pillai approx F num Df den Df Pr(>F)
Condition                 1 0.020512  1.16227      4    222 0.3284
Teacher                   1 0.001475  0.08201      4    222 0.9878
Level                     3 0.057023  1.08504     12    672 0.3696
Condition:Teacher         1 0.019761  1.11886      4    222 0.3484
Condition:Level           3 0.049028  0.93039     12    672 0.5157
Teacher:Level             1 0.020735  1.17519      4    222 0.3226
Condition:Teacher:Level   1 0.012625  0.70967      4    222 0.5861
Residuals               225
```

Based on this analysis, no significant changes were observed. Overall, students did not change in their course interests, and there was no interaction based on Condition, Teacher, or Level. There exist, of course, interactions between students' initial or final interests and their Level, but no interactions exist with regard to the change in course interest over the course of the intervention.

Perception of MILA & the Tutors

As part of the final attitudinal survey, students were also asked to rate their perception of MILA. Perception of MILA was rated according to students' responses to three prompts: "I feel more confident with science now than I did before using MILA.", "I feel more interested in science now than I did before using MILA.", and "MILA was a useful environment for creating scientific models." Each prompt was addressed on a five-point scale, from Strongly Disagree (1 point) to Strongly Agree (5 points). Therefore, students' overall perception of MILA was rated on a scale of 3 to 15.

The overall average score given by students was 9.84 out of 15; the 95% confidence interval around the perception score was (9.52, 10.16). Given that a 9 would be a neutral score (a response of '3' to all three questions, this reflected a slightly positive overall perception of MILA. Following up this overall report, a multivariate analysis of variance was run on students' reported perceptions of MILA along with three independent variables: Condition, Teacher, and Level. The results of this analysis are below.

Table 39: Univariate ANOVA results on students' reported perceptions of MILA. Statistically significant interactions are observed based on Teacher and Condition, with Teacher A's students perceiving MILA more highly than Teacher B's and Control group students perceiving MILA more highly than Experimental group students.

```
Response MILA.A :
                         Df  Sum Sq Mean Sq F value    Pr(>F)
Condition                 1   36.51  36.505  5.9457 0.0155218
Teacher                   1   71.80  71.798 11.6939 0.0007433
Level                     2    1.17   0.586  0.0954 0.9090079
Condition:Teacher         1    6.07   6.073  0.9891 0.3210195
Condition:Level           2    5.25   2.627  0.4278 0.6524654
Teacher:Level             1    7.02   7.016  1.1426 0.2862287
Condition:Teacher:Level   1    0.35   0.350  0.0570 0.8114788
Residuals               227 1393.74   6.140
```

This analysis revealed interactions between students' perceptions of MILA and two independent variables, Condition and Teacher. With regard to Condition, students in the Control condition rated MILA more highly than students in the Experimental condition, 10.30 (σ = 2.40) to 9.51 (σ = 2.59). The 95% confidence interval for the perception of MILA for the Experimental group is (9.08, 9.94), while the interval for the Control group is (9.83, 10.77), suggesting that while both groups had a slightly positive perception of MILA, the Control group's perception was more positive. Students in Teacher A's classroom reported a perception of MILA of 10.45 (σ = 2.48) while students in Teacher B's classroom reported 9.34 (σ = 2.49), showing that Teacher A's students had a more positive perception of MILA than Teacher B's. The full summary of students' perceptions of MILA is provided in Table 40, below. It is worth noting that an alternative explanation of the improvement in the Attitudes toward Scientific Inquiry and Career Interest in Science among the Experimental students may have otherwise been described as a side effect of a positive perception of the tools themselves; however, the presence of a slightly superior opinion of MILA among students in the Control group suggests that this cannot explain the attitudinal improvements documented previously.

Table 40: Perception of MILA for each independent variable and combination of independent variables. Significant interactions were observed based on Condition and Teacher. See Table 22 for further information about the groups of students summarized in each cell.

Students' Perceptions of MILA

<table>
<tr><td colspan="3">Teacher A</td><td colspan="3">Teacher B</td><td colspan="3">Teacher A</td><td colspan="3">Teacher B</td></tr>
<tr><td colspan="3">10.71
(2.22)
n = 45</td><td colspan="3">9.96
(2.50)
n = 54</td><td colspan="3">10.26
(2.65)
n = 62</td><td colspan="3">8.89
(2.40)
n = 76</td></tr>
<tr><td colspan="6">Control</td><td colspan="6">Experimental</td></tr>
<tr><td colspan="6">10.30
(2.40)
n = 99</td><td colspan="6">9.51
(2.59)
n = 138</td></tr>
<tr><td colspan="2">On-Level</td><td colspan="2">Gifted</td><td colspan="2">Sp. Needs</td><td colspan="2">On-Level</td><td colspan="2">Gifted</td><td colspan="2">Sp. Needs</td></tr>
<tr><td colspan="2">10.28
(2.37)
n = 71</td><td colspan="2">10.83
(1.98)
n = 18</td><td colspan="2">9.50
(3.21)
n = 10</td><td colspan="2">9.35
(2.53)
n = 113</td><td colspan="2">10.19
(2.69)
n = 16</td><td colspan="2">10.33
(3.20)
n = 9</td></tr>
<tr><td>A</td><td>B</td><td>A</td><td>B</td><td>A</td><td>B</td><td>A</td><td>B</td><td>A</td><td>B</td><td>A</td><td>B</td></tr>
<tr><td>10.91
(1.90)
n = 23</td><td>9.98
(2.52)
n = 48</td><td>10.83
(1.98)
n = 18</td><td>--
--
n = 0</td><td>9.00
(4.40)
n = 4</td><td>9.83
(2.56)
n = 6</td><td>10.26
(2.53)
n = 38</td><td>8.88
(2.41)
n = 75</td><td>10.19
(2.69)
n = 16</td><td>--
--
n = 0</td><td>10.38
(3.42)
n = 8</td><td>10.00
--
n = 1</td></tr>
</table>

Students in the Experimental condition were also surveyed on their perception of the tutoring system. Students evaluated MILA–T across ten prompts, provided in Appendix A as part of the attitudinal survey. Each prompt was evaluated on a scale of 1 (Strongly Disagree) to 5 (Strongly Agree), for an overall scale of 10 to 50 with a neutral score of '30'. The average student response was 32.80, and the 95% confidence interval surrounding the average response was (31.53, 34.07), reflecting a slightly positive perception of MILA–T.

Table 41: Univariate ANOVA results on students' reported perceptions of MILA–T. A statistically significant interaction is observed based on Teacher, with Teacher A's students perceiving MILA–T more highly than Teacher B's.

```
Response Tutors.A :
               Df Sum Sq Mean Sq F value    Pr(>F)
Teacher         1  668.9  668.93 12.2835 0.0006233
Level           2   81.4   40.68  0.7470 0.4757521
Teacher:Level   1    0.6    0.64  0.0118 0.9136102
Residuals     133 7242.8   54.46
```

A univariate analysis of variance was run on the results of students' perceptions of the tutoring system. Because only students in the Experimental group had access to the tutoring system, only two independent variables were used: Teacher and Student Level. The results of this analysis are provided in Table 41 above.

Table 42: Perception of MILA for each independent variable and combination of independent variables. Significant interactions were observed based on Condition and Teacher. See Table 22 for further information about the groups of students summarized in each cell (note that this table contains only the Experimental half of the table in Table 22).

Experimental Students' Perceptions of MILA–T

<table>
<tr><td colspan="3">Teacher A</td><td colspan="3">Teacher B</td></tr>
<tr><td colspan="3">35.24
(6.91)
n = 62</td><td colspan="3">30.82
(7.67)
n = 76</td></tr>
<tr><td colspan="6">Experimental</td></tr>
<tr><td colspan="6">32.80
(7.64)
n = 138</td></tr>
<tr><td colspan="2">On-Level</td><td colspan="2">Gifted</td><td colspan="2">Sp. Needs</td></tr>
<tr><td colspan="2">32.01
(7.59)
n = 113</td><td colspan="2">36.06
(7.06)
n = 16</td><td colspan="2">37.00
(6.98)
n = 9</td></tr>
<tr><td>A</td><td>B</td><td>A</td><td>B</td><td>A</td><td>B</td></tr>
<tr><td>34.42
(6.80)
n = 38</td><td>30.79
(7.72)
n = 75</td><td>36.06
(7.06)
n = 16</td><td>--
--
n = 0</td><td>37.50
(7.29)
n = 8</td><td>33.00
--
n = 1</td></tr>
</table>

Based on this analysis, a statistically significant interaction between the teacher and the perception of MILA–T is observed. Students in Teacher A's classes perceive MILA–T more highly than students in Teacher B's. While there appears to also be an interaction with students' level (Gifted and Special Needs students report their perceptions of MILA–T at 36.06 and 37.00 respectively, while On-Level students report their perception of MILA–T at 32.01), this interaction is actually explained by the interaction with the teacher variable and the fact that all gifted students are in Teacher A's classrooms. The full results of this analysis are provided above in Table 42.

Summary of Additional Results

Although these analyses do not directly address the hypotheses at the beginning of this dissertation, they help provide a more complete picture of the entire classroom intervention. However, there also exist confounds that limit the reliability of these data. First, students rating the tools were aware that this work was important to the guests in the classroom, and thus quite likely rated it more highly as a friendly gesture. Second and more importantly, however, certain elements of MILA and MILA–T are intentionally structured to challenge students to perform better, to critique students' past performance, and to set constraints and standards for the type of results students are expected to deliver. All of these things serve the learning goals of the intervention, but also risk reducing students' positive perception of the tools. Thus, while the modestly positive perceptions of MILA and MILA–T observed in the survey results are encouraging, the learning gains remain the most important positive takeaway of this study.

CHAPTER 5

TEAM RESULTS

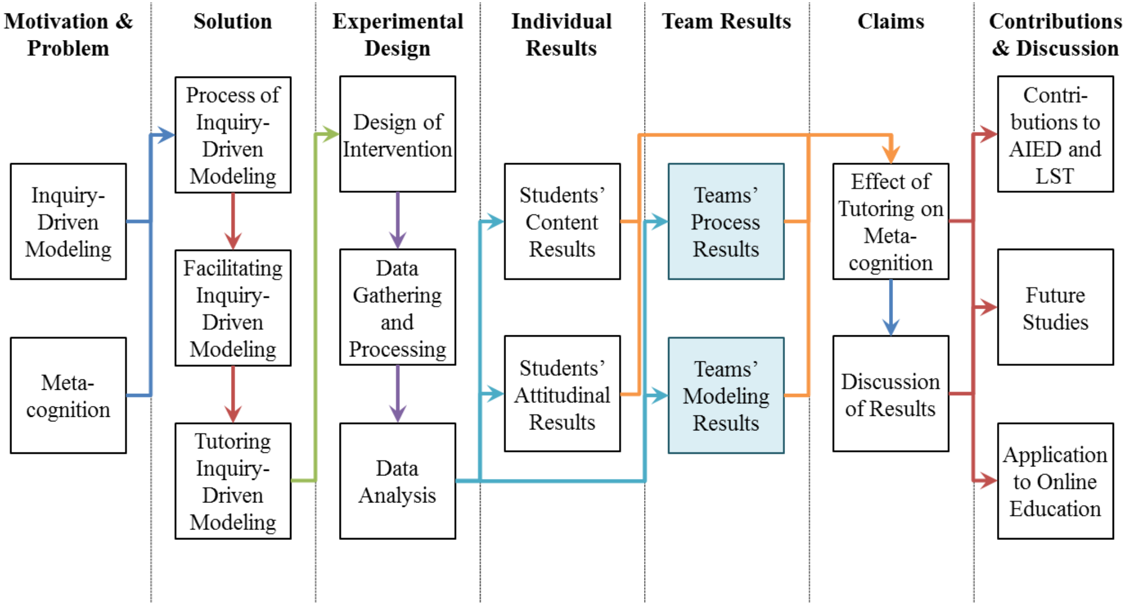

Motivation & Problem
Solution
Experimental Design
Individual Results
Team Results
Claims
Contributions & Discussion
Inquiry-Driven Modeling
Meta-cognition
Process of Inquiry-Driven Modeling
Facilitating Inquiry-Driven Modeling
Tutoring Inquiry-Driven Modeling
Design of Intervention
Data Gathering and Processing
Data Analysis
Students' Content Results
Students' Attitudinal Results
Teams' Process Results
Teams' Modeling Results
Effect of Tutoring on Meta-cognition
Discussion of Results
Contri-butions to AIED and LST
Future Studies
Application to Online Education

Figure 24: The fifth chapter of this dissertation covers analysis at the team level. First, it looks at analysis of the process in which teams of students engage, and second it looks at the models that those teams generate.

Chapter 4 of this dissertation explored the results at the level of individual students, including individuals' understanding of inquiry-driven modeling, dispositions toward science, scientific inquiry, and careers in science, and dispositions toward the intervention as a whole. Chapter 5 shifts to focus on analysis of teams. Specifically, Chapter 5 examines the process of model construction and the resultant explanations that teams of students generate. Chapter 5 builds on the Preliminary Analysis conducted in Chapter 4, which revealed no statistically significant differences between students in the Control and Experimental groups. Thus, systematic differences between the performance of Control and Experimental teams can be reasonably attributed to the presence of the tutoring system. This chapter will first explore the third hypothesis, examining process of model construction that teams of students have undergone. It will then explore the fourth hypothesis, examining the actual explanations that teams generate. Finally, it will also report on interviews with the teachers at the conclusion of the intervention, summarizing their perception of the results in the classroom.

Analyzing teams in this context presents a layer of abstraction between individual understanding and observable results; it is not always clear how the products produced by a team reflect on the understanding of each individual learner in the team. Oftentimes, superior performance of one team over another may simply reflect a single superior team member. The aim of this controlled experiment, however, is to evaluate the usefulness of MILA–T in improving performance overall; assuming that a team cannot improve without at least one of its individuals improving, this experiment will still reveal whether MILA–T is helpful to a team of collaborative learners. Although this precludes claims about individual learners, the forthcoming conclusion that teams do improve as a result of engagement with MILA–T is nonetheless indicative of a desirable overall effect that may be symptomatic of, but does not necessarily demonstrate, improved metacognition. More importantly, although examining teams limits the measurability of individual learners, a collaborative approach to an inquiry activity is a more authentic representation of the way in which science is conducted in the real world (Bray 2000; Wellman & Lipton 2004), and literature on science education similarly supports the need to facilitate collaborative inquiry at the expense of measurability (Suthers, Toth & Weiner 1997; Manlove, Lazonder & Jong 2006).

Hypothesis #3: Procedural Execution

The third hypothesis of this work hypothesized that as a result of engagement with a metacognitive tutoring system during interaction with MILA, teams would better execute the process of inquiry-driven modeling. The third hypothesis is broken down into two sub-hypotheses: first, that teams will better engage in inquiry-driven modeling while receiving feedback from a metacognitive tutoring system, and second, that teams who have previously participated in a modeling activity with a metacognitive tutoring system will better engage in inquiry-driven modeling even when the tutoring system is no longer in use. The first hypothesis reflects the improvement in modeling while receiving feedback, while the second hypothesis reflects the transferal of that improvement to a new project even when such feedback is no longer being received.

The process of inquiry-driven modeling itself, as developed based on the literature on scientific modeling and scientific inquiry, was described previously in Chapter 2. Many facets of the process are contained within the minds of the students in the team or in the conversations amongst the team members during the modeling process, making analysis difficult; ideally, teams would develop the capacity to run through the cycle several times each minute, consistently engaging in a process of considering new information, reflecting on the current understanding, modifying it accordingly, and returning to gathering new information.

Within the software, however, a number of tasks are tracked to allow for insight into the process of model construction in which teams engage. While much of the process of model construction happens within the minds of the teammates and in their conversations, portions of it are also visible in the software; in this context, all model construction and revision actions are logged, providing information on the construction of an explicit, shared model within MILA. This information, then, supports an examination of the major instances of model construction and revision. These analyses can also be found in Joyner & Goel 2015.

Action Categorization

Actions within the software were initially logged at a low level of granularity, such as adding or moving a particular node or changing the category for a single piece of evidence. In order to facilitate analysis of higher-order behaviors along the lines of model construction and revision, these low-level activities were first categorized into a schema reflecting the higher-order activities that each suggested. These higher-order activities were derived from the process of inquiry-driven modeling itself in order to analyze the data in terms of the learning goal for the exercise. The activities derived are: Start, Writing Problem Definition, Proposing Hypothesis, Constructing Model, Constructing Evidence, Revising Model, Revising Evidence, Dismissing Model, Reconsidering Model, Taking Notes, Using Simulation, Switching Model, and Consulting Tutor (for teams in the Experimental group during the Learning project). These activities were written to correspond to the process of inquiry-driven modeling, and roughly map to the previous model as shown in Figure 25 below.

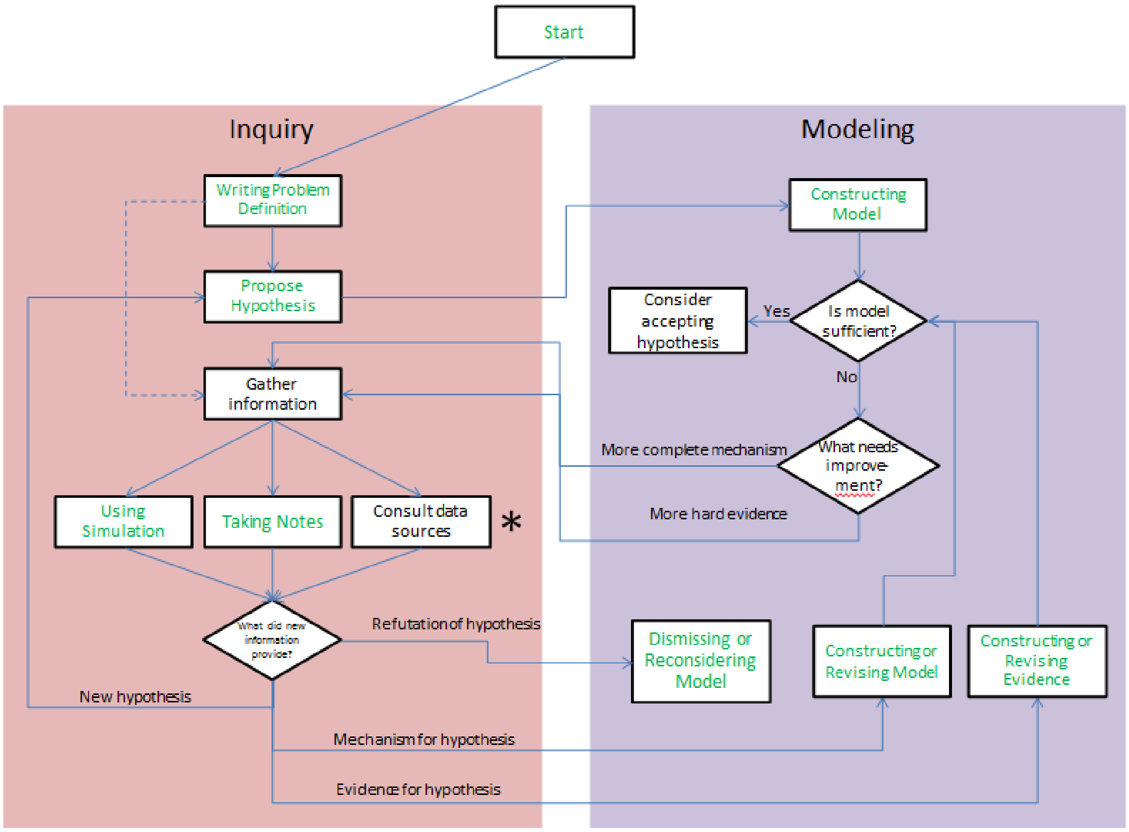


Figure 25: The inquiry-driven modeling process described earlier. Written in green are the tasks and activities identified in teams' model construction processes, mapped onto the overall process of inquiry-driven modeling.

Certain activities are present in multiple places, such as model construction either before any investigation has taken place or after gathering more information. Other activities in which teams participate are not tracked within the software; for example, teams may also gather information from in-class activities, in-class experts, or other sources outside the software, unavailable here. One task, Switching Model, is not contained within this model of the inquiry-driven modeling process; the Switching Model task occurs when a team switches what model it is currently constructing or revising, and therefore is a task at a greater level of abstraction than the model shown here. One can imagine the inquiry-driven modeling task being carried out for each hypothesis and model individually.

Table 43 below shows the mapping of each low-level action within MILA to a higher-level activity in the inquiry-driven modeling process. Under this categorization, model construction is defined as instances wherein something new is inserted into the model, and evidence construction is defined as instances wherein a new piece of evidence is added to the model. In contrast, model revision is defined as instances where a previously-existing portion of the model is modified or removed, and evidence revision is defined as instances where a previously-added bit of evidence is modified or removed.

A number of tracked actions within MILA are not contained within this categorization because they were not regarded to be strongly indicative of a particular phase of the inquiry-driven modeling process. These actions are: Component Moved, Evidence Viewed, Project Opened, Switching Model, Tutor Switched, Tutor Text Feedback Given, Tutor Hidden, Tutor Question Clicked. The four tutor-specific actions are coincident with the Tutor Shown action recorded previously, while Component Moved, Switching Model, Evidence Viewed, and Project Opened are not individually indicative of any activity without an accompanying other action such as Component Removed or Evidence Changed.

Markov Chains

After deriving the categorization of actions into activities and the correspondence of those activities to the inquiry-driven modeling process, the activities were compiled together into Markov Chains (Kemeny & Snell 1960) reflecting the transition amongst activities among all teams. As mentioned previously, Markov chains were chosen instead of hidden Markov models because of the desire to map and compare teams' behaviors to a preexisting model of the target process or skill rather than develop a grounded description of teams' behaviors (Ghahramani 2001). A significant limitation of analysis based on Markov chains is the assumption that the likelihood of a subsequent state is based solely on the current state rather than anything about the history of state transitions preceding the current state (Lopez, Hermanns, & Katoen 2001). In reality, we would likely assume that determination of the next activity that teams enter would be based not only on the previous activity, but on the history of activities prior to the previous activity as well. However, given the size of the dataset (22,547 individual actions), methods of analysis that include a broader activity history would be computationally intractable, and so this analysis builds on the Markov chains proposed here.

Table 43: The compilation of certain low-level actions within MILA to higher-order activities in the inquiry-driven modeling process.

Activity	Actions
Start	Start
Writing Problem Definition	Problem Definition Changed
Proposing Hypothesis	New Model Added
Constructing Model	Component Added, New Connection
Revising Model	Component Edited, Component Removed, Component Direction Changed, Removed Edge, Connection Direction Toggled
Constructing Evidence	Evidence Added
Revising Evidence	Evidence Changed, Evidence Removed
Dismissing Model	Model Dismissed
Reconsidering Model	Model Reconsidered
Taking Notes	Opened Notes Dialog
Using Simulation	Opened Simulation Dialog
Consulting Tutor	Tutor Feedback Received, Tutor Shown

As with the data from the individual analyses of students' performance on the content test and attitudinal metrics, individual Markov chains could be constructed based on a variety of subdivisions of the data, including by Teacher, by Class Level, and by Condition. This analysis will primarily consider dividing the data into four chains: the Control condition on the Learning project, the Control condition on the Transfer project, the Experimental condition on the Learning project, and the Experimental condition on the Transfer project. Comparisons will focus on comparing across conditions within projects.

Due to the complexity of these Markov chains (featuring transitions amongst twelve different tasks), these chains are presented here as tables. Each cell of the chart describes the percentage of the time that a team transitioned from the activity for that row to the activity for that column; the length of the green bar in the cell summarizes the value within that cell. Five chains are presented: one chain for each combination of Condition and Project, and an additional chain for the Experimental group's Learning project that excludes the Consulting Tutor task in order to draw more direct comparisons with the Control condition's Learning project.

Analysis of these Markov chains takes three forms. First, a repeated Z-test is used to test for targeted improvements in certain dimensions of the inquiry-driven modeling process between comparable Markov chains in the Control and Experimental groups. Second, a repeated χ^2 test is run on corresponding portions of comparable chains to test for the presence of any difference between the chains for the Control and Experimental groups. Third, a targeted qualitative analysis specifically on the Markov chains for the Experimental group is used to identify the precise placement and role of the tutoring system within the broader inquiry-driven modeling process.

Experimental Group, Learning Project, Including Interactions with MILA–T

	Number of Instances	Constructing Evidence	Constructing Model	Consulting Tutor	Dismissing Model	Proposing Hypothesis	Reconsidering Model	Revising Evidence	Revising Model	Taking Notes	Using Simulation	Writing Problem Definition
Constructing Evidence	866	51.96%	3.35%	23.09%	1.62%	0.35%	0.00%	7.39%	8.31%	1.62%	2.08%	0.23%
Constructing Model	1520	5.07%	55.59%	11.91%	0.39%	2.30%	0.13%	0.33%	23.16%	0.33%	0.53%	0.26%
Consulting Tutor	1821	8.68%	11.20%	56.34%	2.20%	2.25%	0.55%	2.14%	7.80%	2.31%	4.01%	2.53%
Dismissing Model	167	1.20%	5.39%	32.34%	21.56%	5.99%	23.95%	1.80%	4.79%	1.80%	0.60%	0.60%
Proposing Hypothesis	194	0.00%	41.24%	24.23%	8.76%	19.59%	0.00%	0.00%	2.58%	1.03%	0.52%	2.06%
Reconsidering Model	54	0.00%	5.56%	27.78%	51.85%	5.56%	1.85%	0.00%	3.70%	0.00%	3.70%	0.00%
Revising Evidence	229	27.51%	2.62%	17.03%	0.44%	0.44%	0.00%	41.48%	7.86%	1.75%	0.44%	0.44%
Revising Model	2100	4.90%	15.24%	5.86%	0.76%	0.62%	0.05%	1.10%	70.62%	0.33%	0.33%	0.19%
Start	50	0.00%	0.00%	30.00%	0.00%	8.00%	0.00%	0.00%	0.00%	0.00%	0.00%	62.00%
Taking Notes	102	8.82%	5.88%	36.27%	3.92%	0.00%	0.00%	0.98%	2.94%	37.25%	2.94%	0.98%
Using Simulation	242	2.89%	4.13%	26.03%	2.07%	2.48%	0.00%	0.41%	6.61%	2.48%	52.48%	0.41%
Writing Problem Definition	110	1.82%	7.27%	33.64%	0.00%	38.18%	0.00%	0.00%	2.73%	0.91%	0.91%	14.55%

Markov Chain 1: Patterns of interaction with MILA in the Experimental group during the Learning project. Each cell represents the percentage of the instances of the activity for the row that were followed by the activity for the column. For example, the first first row (Constructing Evidence) and the first column (Constructing Evidence) show that 51.96% of the time that students construct evidence, they follow by constructing evidence again. The cell below shows that 5.07% of the time that teams construct their model, they follow by constructing evidence. In this chain, instances of consultation with MILA–T are included.

Experimental Group, Learning Project, Not Including Interactions with MILA–T

	Number of Instances	Constructing Evidence	Constructing Model	Dismissing Model	Proposing Hypothesis	Reconsidering Model	Revising Evidence	Revising Model	Taking Notes	Using Simulation	Writing Problem Definition
Constructing Evidence	866	61.65%	5.79%	2.55%	0.46%	0.00%	10.31%	12.05%	2.67%	3.82%	0.70%
Constructing Model	1520	7.02%	60.72%	0.59%	2.82%	0.20%	0.39%	26.10%	0.46%	0.98%	0.72%
Dismissing Model	167	2.98%	13.69%	26.79%	8.93%	27.98%	3.57%	5.95%	7.14%	1.19%	1.79%
Proposing Hypothesis	194	0.00%	55.61%	9.18%	25.00%	0.00%	0.00%	4.08%	1.53%	0.51%	4.08%
Reconsidering Model	54	1.85%	12.96%	62.96%	9.26%	1.85%	0.00%	3.70%	0.00%	5.56%	1.85%
Revising Evidence	229	32.02%	5.26%	0.88%	0.88%	0.44%	43.42%	11.84%	3.95%	0.88%	0.44%
Revising Model	2100	5.99%	16.89%	1.09%	0.86%	0.10%	1.09%	72.55%	0.48%	0.67%	0.29%
Start	50	0.00%	0.00%	0.00%	8.00%	0.00%	0.00%	0.00%	2.00%	0.00%	90.00%
Taking Notes	102	12.00%	10.00%	7.00%	3.00%	0.00%	4.00%	6.00%	48.00%	7.00%	3.00%
Using Simulation	242	4.96%	8.26%	2.89%	2.89%	0.00%	1.24%	8.68%	2.89%	67.77%	0.41%
Writing Problem Definition	110	2.75%	11.93%	0.92%	45.87%	0.00%	0.92%	6.42%	2.75%	1.83%	26.61%

Markov Chain 2: Patterns of interaction with MILA in the Experimental group during the Learning project. Each cell represents the percentage of the instances of the activity for the row that were followed by the activity for the column. For example, the first first row (Constructing Evidence) and the first column (Constructing Evidence) show that 61.65% of the time that teams construct evidence, they follow by constructing evidence again. The cell below shows that 7.02% of the time that teams construct their model, they follow by constructing evidence. In this chain, instances of consultation with MILA–T are not included.

Control Group, Learning Project

	Number of Instances	Constructing Evidence	Constructing Model	Dismissing Model	Proposing Hypothesis	Reconsidering Model	Revising Evidence	Revising Model	Taking Notes	Using Simulation	Writing Problem Definition
Constructing Evidence	511	60.08%	5.48%	1.76%	0.78%	0.00%	7.83%	15.07%	3.91%	4.89%	0.20%
Constructing Model	1063	4.51%	62.88%	1.13%	3.29%	0.00%	0.28%	26.22%	0.56%	0.85%	0.28%
Dismissing Model	123	3.25%	6.50%	24.39%	18.70%	30.89%	0.81%	9.76%	4.88%	0.00%	0.81%
Proposing Hypothesis	138	0.72%	70.29%	7.97%	18.12%	1.45%	0.00%	0.72%	0.72%	0.00%	0.00%
Reconsidering Model	45	4.44%	4.44%	64.44%	11.11%	4.44%	0.00%	4.44%	2.22%	0.00%	4.44%
Revising Evidence	111	31.25%	3.57%	0.89%	0.89%	0.00%	43.75%	12.50%	2.68%	3.57%	0.89%
Revising Model	1229	7.24%	19.45%	1.95%	0.81%	0.08%	1.22%	66.97%	0.65%	1.38%	0.24%
Start	34	0.00%	0.00%	0.00%	11.76%	0.00%	0.00%	0.00%	0.00%	0.00%	88.24%
Taking Notes	78	12.20%	4.88%	3.66%	4.88%	1.22%	2.44%	6.10%	57.32%	7.32%	0.00%
Using Simulation	165	10.24%	7.23%	2.41%	3.01%	0.00%	1.20%	9.04%	3.01%	63.86%	0.00%
Writing Problem Definition	44	2.27%	9.09%	0.00%	68.18%	2.27%	0.00%	4.55%	6.82%	0.00%	6.82%

Markov Chain 3: Patterns of interaction with MILA in the Control group during the Learning project. Each cell represents the percentage of the instances of the activity for the row that were followed by the activity for the column. For example, the first first row (Constructing Evidence) and the first column (Constructing Evidence) show that 60.08% of the time that teams construct evidence, they follow by constructing evidence again. The cell below shows that 4.51% of the time that teams construct their model, they follow by constructing evidence.

Experimental Group, Transfer Project

	Number of Instances	Constructing Evidence	Constructing Model	Dismissing Model	Proposing Hypothesis	Reconsidering Model	Revising Evidence	Revising Model	Taking Notes	Using Simulation	Writing Problem Definition
Constructing Evidence	366	65.30%	5.74%	1.09%	0.27%	0.00%	9.84%	12.30%	3.28%	2.19%	0.00%
Constructing Model	710	6.90%	64.08%	0.42%	0.56%	0.14%	0.28%	25.77%	0.14%	1.55%	0.14%
Dismissing Model	32	3.13%	9.38%	21.88%	28.13%	15.63%	0.00%	6.25%	6.25%	3.13%	6.25%
Proposing Hypothesis	75	0.00%	68.00%	12.00%	10.67%	1.33%	0.00%	0.00%	0.00%	5.33%	2.67%
Reconsidering Model	9	0.00%	44.44%	33.33%	11.11%	11.11%	0.00%	0.00%	0.00%	0.00%	0.00%
Revising Evidence	60	51.67%	3.33%	0.00%	0.00%	0.00%	25.00%	15.00%	3.33%	1.67%	0.00%
Revising Model	646	8.36%	25.70%	0.62%	0.15%	0.00%	0.93%	62.23%	0.62%	1.39%	0.00%
Start	47	0.00%	0.00%	0.00%	25.53%	0.00%	0.00%	0.00%	0.00%	10.64%	63.83%
Taking Notes	22	4.55%	4.55%	9.09%	13.64%	0.00%	9.09%	13.64%	40.91%	4.55%	0.00%
Using Simulation	53	22.64%	18.87%	0.00%	5.66%	1.89%	3.77%	15.09%	1.89%	22.64%	7.55%
Writing Problem Definition	44	2.27%	2.27%	0.00%	77.27%	0.00%	0.00%	0.00%	0.00%	4.55%	13.64%

Markov Chain 4: Patterns of interaction with MILA in the Experimental group during the Transfer project. Each cell represents the percentage of the instances of the activity for the row that were followed by the activity for the column. For example, the first first row (Constructing Evidence) and the first column (Constructing Evidence) show that 65.30% of the time that teams construct evidence, they follow by constructing evidence again. The cell below shows that 6.90% of the time that teams construct their model, they follow by constructing evidence.

Control Group, Transfer Project

	Number of Instances	Constructing Evidence	Constructing Model	Dismissing Model	Proposing Hypothesis	Reconsidering Model	Revising Evidence	Revising Model	Taking Notes	Using Simulation	Writing Problem Definition
Constructing Evidence	245	68.57%	4.08%	0.41%	1.22%	0.00%	11.43%	8.57%	0.41%	4.08%	1.22%
Constructing Model	435	9.20%	63.22%	0.00%	0.46%	0.00%	0.23%	24.83%	0.00%	1.61%	0.46%
Dismissing Model	23	0.00%	13.04%	4.35%	30.43%	26.09%	0.00%	0.00%	4.35%	13.04%	8.70%
Proposing Hypothesis	60	3.33%	51.67%	18.33%	10.00%	5.00%	0.00%	0.00%	0.00%	5.00%	6.67%
Reconsidering Model	10	10.00%	20.00%	30.00%	20.00%	0.00%	0.00%	0.00%	0.00%	10.00%	10.00%
Revising Evidence	44	45.45%	6.82%	2.27%	0.00%	0.00%	31.82%	9.09%	0.00%	4.55%	0.00%
Revising Model	350	6.00%	27.43%	0.57%	0.86%	0.00%	0.86%	61.14%	0.00%	1.71%	1.43%
Start	34	0.00%	0.00%	0.00%	38.24%	0.00%	0.00%	0.00%	0.00%	17.65%	44.12%
Taking Notes	6	0.00%	16.67%	0.00%	16.67%	0.00%	0.00%	16.67%	0.00%	50.00%	0.00%
Using Simulation	54	16.67%	24.07%	3.70%	9.26%	1.85%	3.70%	5.56%	1.85%	24.07%	9.26%
Writing Problem Definition	41	4.88%	7.32%	4.88%	48.78%	0.00%	0.00%	7.32%	0.00%	12.20%	14.63%

Markov Chain 5: Patterns of interaction with MILA in the Control group during the Transfer project. Each cell represents the percentage of the instances of the activity for the row that were followed by the activity for the column. For example, the first first row (Constructing Evidence) and the first column (Constructing Evidence) show that 68.57% of the time that teams construct evidence, they follow by constructing evidence again. The cell below shows that 9.20% of the time that teams construct their model, they follow by constructing evidence.

Targeted Improvement in Inquiry-Driven Modeling

Prior to calculating the values of the Markov chains for each combination of Condition and Project, nine desired changes potentially identifiable by these Markov chains were identified. These nine changes represented differences between the Control and the Experimental groups that the planned Markov chains would be capable of identifying. Table 44 below identifies the differences and the measurements from the Markov chains that would signify the presence of the difference. Broadly, the improvements are designed to reflect perceived weaknesses in teams of students in past interventions in the project history of ACT (Goel et al. 2013), EMT (Joyner et al. 2012), and MILA. In past interventions, teams exhibited a handful of weaknesses and errors in inquiry-driven modeling, identified previously in Figure 5 on page 29. For example, teams often:

- Jump straight to proposing hypotheses before establishing a strong definition of the phenomenon that they were trying to explain. This leads to effects like teams becoming distracted by superficial observations (such as the smell of Lake Clara Meer driving people away) that are not part of the causal chain leading to the phenomenon.
- Rely on asking the teacher for the final correct explanation rather than exploring themselves. This reflects both a tendency to deemphasize self-driven inquiry as well as an underlying misconception that science is about finding an established correct answer rather than developing well-defended arguments and explanations.
- Attempt to fit multiple hypotheses or new hypotheses into a single model rather than developing models that targeted explaining specific hypotheses. This leads to teams' models reflecting a general speculation on everything going on in the system rather than a targeted explanation of the relationship between the hypothesized cause and the phenomenon.

- Making up information in support of their hypotheses rather than inquiring into the data or literature. In this sense, making up information is not malicious, but rather reflects teams' tendencies to design evidential justifications or mechanistic explanations that make sense in their own minds rather than being grounded in the literature or investigation of the phenomenon. This tendency is addressed more strongly in Hypothesis #4's emphasis on certain kinds of evidence over these sorts of logical explanations.
- Accepting a hypothesis too soon rather than fleshing it out sufficiently. Embedded in this error is the tendency of teams to focus solely on putting information into their models rather than modifying and revising the information they have previously provided. In earlier studies, teams tended to construct models that made sense to them and discuss how to justify the models later, which led to a hesitance to ever remove claims or connections made in previous models. Instead, multiple iterations through the model revision process are desired before ultimately landing on a hypothesis teams are willing to submit.
- Using simulations in an undirected and unintentional manner. Teams often become distracted by the game-like elements of simulations and simply "play" with them instead of controlling variables and looking for specific relationships.

MILA–T was, in many ways, designed to address precisely these mistakes, as described earlier. In order to connect these learning goals with MILA–T, the desired differences based on the presence of MILA–T are also derived from these errors and the broader inquiry-driven modeling process. In analyzing the Markov chains, positive changes would be instances wherein teams demonstrated greater tendencies to articulate their problem, revise prior claims and justifications based on new evidence, and dismiss hypotheses for which there is insufficient support. Of course, not all activities that teams complete as part of the inquiry-driven modeling process are contained within the software; while teams were encouraged to use the note-taking facilities of MILA to organize information from articles and in-class experiments, teams did jump straight into construction and revision. For this reason, the desired differences examined in this analysis all take the form of increased presence of a desirable actions rather than decreased presence of an undesirable one, given that identification of an undesirable activity would potentially misrepresent activity in the classroom. For example, we might witness a team propose a hypothesis, construct a model, and then immediately dismiss the model. It would be presumptuous to treat this as an error of jumping straight from construction to dismissal without gathering information because an activity outside of the software, such as reading a source provided in class, might provide the information necessary to choose to dismiss that hypothesis. Thus, rather than infer errors, we instead look for specific desirable transitions that would serve as positive indicators of execution of the desirable inquiry-driven modeling process.

Along these lines, nine desirable differences were identified; this list is not exhaustive and would not be sufficient to determine that improvement did not occur, but would be capable of demonstrating that some improvement did occur. These differences were identified prior to development of the Markov chains summarizing the results, and thus were developed independently of knowledge of the patterns of interaction in the Control and Experimental groups. Table 44 below summarizes the nature of these desired differences and the operationalized potential indicators of the presence of these differences within the Markov chains. The name given at the left provides the identifier that will be used when testing for the presence of the given difference in the subsequent analysis.

Table 44: The desired differences in inquiry-driven modeling based on interaction with MILA–T.

Name	Difference	Difference in Markov Chains
#1	Greater incidence of articulating the problem prior to beginning any modeling or inquiry activities.	There is a greater probability of transitioning from Start to Writing Problem Definition.
#2	Greater incidence of using the results of data-gathering activities to construct new evidential justifications for a model's claims.	There is a greater probability of transitioning from Taking Notes and Using Simulation to Constructing Evidence.
#3	Greater incidence of using the results of data-gathering activities to revise previous evidential justifications for a model's claims.	There is a greater probability of transitioning from Taking Notes and Using Simulation to Revising Evidence.
#4	Greater incidence of using the results of data-gathering activities to refute previous hypotheses.	There is a greater probability of transitioning from Taking Notes and Using Simulation to Dismissing Model.
#5	Greater incidence of using the results of data-gathering activities to propose new hypotheses.	There is a greater probability of transitioning from Taking Notes and Using Simulation to Proposing Hypothesis.
#6	Greater incidence of revising problem definition throughout the inquiry-driven modeling process.	There is a greater probability of an action being Writing Problem Definition even after removing instances of Writing Problem Definition following from Start.
#7	Greater overall incidence of model revision.	There is a greater probability of an action being Revising Model.
#8	Greater incidence of model construction activities spawning further information-gathering activities.	There is a greater probability of Taking Notes or Using Simulation following from Model Construction.
#9	Greater overall incidence of note-taking.	There is a greater probability of an action being Taking Notes.

Repeated Z-test Analysis

These nine desired differences from the Control group based on the Experimental group's interaction with MILA–T represent nine process hypotheses as to potential improvements that MILA–T may bring. Each process hypothesis is similarly present in both the Learning and Transfer projects. Thus, each of these eighteen process hypotheses must be tested. However, given that the anticipated differences are dispersed amongst a much larger data set, these hypotheses are tested using a repeated Z-test for each process hypothesis and project. Given that repeating a Z-test drastically raises the odds of a false positive observation, the results of this analysis are subsequently processed through an additional test to examine the likelihood of the given observations if no true differences were present.

Table 45 gives the results of the two-tailed Z-tests for the nine hypotheses for the Learning project. The column at the far left maps the test to the difference in Table 44. N for the Control and Experimental groups represents the overall number of observed instances of the parent task are observed. For example, the Experimental group has 50 instances of the Start action while the Control group has 34 instances. The % for the Control and Experimental groups represent the number of instances of that action that were followed by the desired action. For example, for process hypothesis #1, 90.0% of the Experimental group's Start activities were followed by the desired Writing Problem Definition action, while 88.24% of the Control group's Start activities were followed by the desired action. For those process hypotheses that refer to a greater overall incidence of an activity (#6, #7, and #9), the measure is taken out of the total number of actions completed. For example, for process hypothesis #7, 37.40% of the Experimental group's actions were in the Revising Model category, while 34.67% of the Control group's actions fell in the same category.

Table 45: Results from a repeated Z-test over the nine hypotheses during the Learning project.

#	Control N	Control Proportion	Experimental N	Experimental Proportion	*p* (z)
#1	34	88.2353%	50	90.0000%	0.795 (0.247)
#2	248	22.4361%	342	16.9587%	0.094 (−1.665)
#3	248	3.6438%	342	5.2397%	0.358 (0.091)
#4	248	6.0682%	342	9.8926%	0.096 (1.663)
#5	248	7.8901%	342	5.8926%	0.338 (−0.956)
#6	3548	0.0039%	5637	0.0122%	0.000 (4.102)
#7	3548	34.6674%	5637	37.3958%	0.008 (2.647)
#8	1064	1.4098%	1525	1.4426%	0.944 (0.069)
#9	3548	2.8185%	5637	2.1820%	0.004 (−2.89)

The repeated Z-test revealed three statistically significant differences. First, teams in the Experimental group displayed a significantly larger propensity to revise their problem definitions over time (#6). Experimental teams performed a total of 69 problem definition revisions beyond writing the initial problem definition over the course of the Learning project compared with 14 revisions for teams in the Control group. Second, Experimental teams demonstrated a significantly higher propensity to revise their models over time (#7), spending 37.4% of their actions on revision compared to 34.7% in the Control group. Third, Control teams actually spent a greater percentage of their actions on note-taking than Experimental teams (#9), spending 2.8% of their actions on note-taking compared to 2.2% for the Experimental group.

As referenced previously, repeating a Z-test raises the odds of a false positive. Thus, these data should be interpreted together rather than in isolation. Taking α = 0.01 as our standard for rejecting a null hypothesis, if none of the hypotheses above were valid, then each would have a 1% (1 in 20) chance of registering as a false positive. A Bernoulli (or binomial) trial reveals a 91.4% chance of no false positives, an 8.3% chance of one false positive, and a 0.3% chance of two or more false positives. Thus, we may infer that no more than one significant difference in the repeated Z-test is a false positive (given that p < 0.01 that more than one false positive would occur). Thus, in this analysis, we may conclude that no more than one of these significant differences was a false positive, and thus, at least two significant differences were observed. We cannot, however, make any determinations as to which difference may have been a false positive. Thus, we may ultimately conclude that **there existed at least one improvement based on participation in the Experimental condition**, and there existed at least one other significant difference as well. Given that it remains statistically possible that the second difference benefited the control group rather than the experimental group, it is difficult to state conclusively that participation in the experimental group improved performance more generally (T.1L), but we may claim it improved performance in at least one particular dimension.

Table 46 gives the results for the same hypotheses on the Transfer project. As with the analysis of the Learning project, analysis of the Transfer project reveals three statistically significant differences. Teams in the Experimental group demonstrate a significantly larger propensity to revise their models over time as well as take notes during the data-gathering process. Teams in the Control group demonstrate a greater likelihood to revise their problem definitions over time. Interestingly, however, there is no statistically significant difference in the overall likelihood of writing a problem definition between the Control and the Experimental groups, suggesting that the increased tendency to revise problem definitions over time in the Control group only serves to equalize such actions relative to the Experimental group's greater number of such actions immediately after beginning the project.

Although a seemingly large difference was seen in the likelihood of Control and Experimental groups to define their problem prior to beginning modeling (process hypothesis #1; 63.8% for the Experimental group, 44.1% for the Control group), the sample size was insufficient to define this difference as statistically significant (*p* = 0.078) (T.1T). Similarly, large differences appear to exist between the Control and Experimental groups' steps after gathering evidence (process hypotheses #2, #3, and #4), but the limited number of instances of data-gathering amongst the groups prevent these differences from being statistically significant.

Table 46: Results from a repeated Z-test over the nine hypotheses during the Transfer project.

#	Control N	Control Proportion	Experimental N	Experimental Proportion	*p* (z)
#1	34	44.1176%	47	63.8298%	0.078 (1.762)
#2	60	16.6667%	75	27.1870%	0.147 (1.454)
#3	60	3.7037%	75	12.8645%	0.061 (1.868)
#4	60	3.7037%	75	9.0909%	0.215 (1.244)
#5	60	25.9259%	75	19.2967%	0.358 (−0.920)
#6	1302	0.0212%	2064	0.0073%	0.001 (−3.511)
#7	1302	27.1889%	2064	31.5891%	0.007 (2.716)
#8	435	1.6092%	710	1.6901%	0.920 (0.104)
#9	1302	0.2304%	2064	1.5019%	0.000 (3.593)

Similar to the analysis of the Learning project, a Bernoulli trial reveals there is less than a 1% chance of drawing more than one false positive. Given three differences, we may thus conclude that a minimum of two of them were not false positives, and thus, we may conclude that **the Experimental group benefited in at least one way**. However, it remains possible that the control group also benefited in one way, and so it is difficult to claim improvement more generally based on participation in the experimental condition (T1.T). We may, however, claim that participation in the experimental condition improved performance along one dimension

Overall Differences in Inquiry-Driven Modeling

The above analysis demonstrated a significant difference in the patterns of interaction and inquiry between the Control group and the Experimental group, but it did not demonstrate that the performance of either group was superior to the other. In addition, it only demonstrated difference between the patterns of inquiry in the two groups according to a narrow set of metrics. In order to more conclusively demonstrate a difference between the activity of the Control and Experimental group, a more thorough analysis covering all the dimensions of the patterns of inquiry is required.

The presentation of the Markov Chains above specifically compares the predicted values based on observations from the Control group to the observed values in the Experimental group. This comparison is performed based on transitions from one action to another. Within the Control group, transitions were observed with certain frequencies; for example, during the Learning project, teams performing a Taking Notes action transitioned into Constructing Model action 4.88% of the time. In the Experimental group during the Learning project, 100 Taking Notes actions were observed. Based on the observation from the Control group, we would expected that if the Control and Experimental groups performed identically, there would be 5 transitions from Taking Notes to Constructing Model (4.88% × 100). Instead, 10 such transitions were observed.

Observed Transitions in the Experimental Group during the Learning Project, and Expected Transitions based on Observations from the Control Group

		Constructing Evidence		Constructing Model		Dismissing Model		Proposing Hypothesis		Reconsidering Model		Revising Evidence		Revising Model		Taking Notes		Using Simulation		Writing Problem Definition	
	Total	*Exp.*	*Obs.*	*Exp.*	*Obs.*	*Exp.*	*Obs.*	*Exp.*	*Obs.*	*Exp.*	*Obs.*	*Exp.*	*Obs.*	*Exp.*	*Obs.*	*Exp.*	*Obs.*	*Exp.*	*Obs.*	*Exp.*	*Obs.*
Constructing Evidence	**863**	518	532	47	50	15	22	7	4	0	0	68	89	130	104	34	23	42	33	2	6
Constructing Model	**1525**	69	107	959	926	17	9	50	43	0	3	4	6	400	398	9	7	13	15	4	11
Dismissing Model	**168**	5	5	11	23	41	45	31	15	52	47	1	6	16	10	8	12	0	2	1	3
Proposing Hypothesis	**196**	1	0	138	109	16	18	36	49	3	0	0	0	1	8	1	3	0	1	0	8
Reconsidering Model	**54**	2	1	2	7	35	34	6	5	2	1	0	0	2	2	1	0	0	3	2	1
Revising Evidence	**228**	71	73	8	12	2	2	2	2	0	1	100	99	29	27	6	9	8	2	2	1
Revising Model	**2102**	152	126	409	355	41	23	17	18	2	2	26	23	1408	1525	14	10	29	14	5	6
Start	**50**	0	0	0	0	0	0	6	4	0	0	0	0	0	0	0	1	0	0	44	45
Taking Notes	**100**	12	12	5	10	4	7	5	3	1	0	2	4	6	6	57	48	7	7	0	3
Using Simulation	**242**	25	12	17	20	6	7	7	7	0	0	3	3	22	21	7	7	155	164	0	1
Writing Problem Definition	**109**	2	3	10	13	0	1	74	50	2	0	0	1	5	7	7	3	0	2	7	29
Total	**9185**	**1332**	**1385**	**2762**	**2592**	**318**	**291**	**378**	**346**	**116**	**99**	**290**	**343**	**3184**	**3338**	**259**	**223**	**432**	**410**	**114**	**158**

Markov Chain 6: Expected and observed values for each transition in the Experimental group during the Learning project. Expected values are derived by multiplying the percent likelihood of a given transition in the Control condition by the total instances of that action in the Experimental condition. For example, in the Control group, 5.48% of Constructing Evidence actions were followed by a Constructing Model action. In the Experimental group, 863 Constructing Evidence actions were observed, meaning that based on the percentage from the Control group, we would expect 47 transitions into Constructing Model actions. In practice, we observed 50 such transitions.

Observed Transitions in the Experimental Group during the Transfer Project, and Expected Transitions based on Observations from the Control Group

		Constructing Evidence		Constructing Model		Dismissing Model		Proposing Hypothesis		Reconsidering Model		Revising Evidence		Revising Model		Taking Notes		Using Simulation		Writing Problem Definition	
	Total	*Exp.*	*Obs.*	*Exp.*	*Obs.*	*Exp.*	*Obs.*	*Exp.*	*Obs.*	*Exp.*	*Obs.*	*Exp.*	*Obs.*	*Exp.*	*Obs.*	*Exp.*	*Obs.*	*Exp.*	*Obs.*	*Exp.*	*Obs.*
Constructing Evidence	**366**	251	239	15	21	1	4	4	1	0	0	42	36	31	45	1	12	15	8	4	0
Constructing Model	**710**	65	49	449	455	0	3	3	4	0	1	2	2	176	183	0	1	11	11	3	1
Dismissing Model	**32**	0	1	4	3	1	7	10	9	8	5	0	0	0	2	1	2	4	1	3	2
Proposing Hypothesis	**75**	3	0	39	51	14	9	8	8	4	1	0	0	0	0	0	0	4	4	5	2
Reconsidering Model	**9**	1	0	2	4	3	3	2	1	0	1	0	0	0	0	0	0	1	0	1	0
Revising Evidence	**60**	27	31	4	2	1	0	0	0	0	0	19	15	5	9	0	2	3	1	0	0
Revising Model	**646**	39	54	177	166	4	4	6	1	0	0	6	6	395	402	0	4	11	9	9	0
Start	**47**	0	0	0	0	0	0	18	12	0	0	0	0	0	0	0	0	8	5	21	30
Taking Notes	**22**	0	1	4	1	0	2	4	3	0	0	0	2	4	3	0	9	11	1	0	0
Using Simulation	**53**	9	12	13	10	2	0	5	3	1	1	2	2	3	8	1	1	13	12	5	4
Writing Problem Definition	**44**	2	1	3	1	2	0	21	34	0	0	0	0	3	0	0	0	5	2	6	6
Total	**2064**	**417**	**388**	**693**	**714**	**36**	**32**	**98**	**76**	**16**	**9**	**76**	**63**	**561**	**652**	**5**	**31**	**94**	**54**	**68**	**45**

Markov Chain 7: Expected and observed values for each transition in the Experimental group during the Learning project. Expected values are derived by multiplying the percent likelihood of a given transition in the Control condition by the total instances of that action in the Experimental condition. For example, in the Control group, 4.08% of Constructing Evidence actions were followed by a Constructing Model action. In the Experimental group, 366 Constructing Evidence actions were observed, meaning that based on the percentage from the Control group, we would expect 15 transitions into Constructing Model actions. In practice, we observed 21 such transitions.

This analysis allows a look at the major differences between the Control and Experimental groups during interaction with MILA on both the Learning and the Transfer projects. The frequency of expected values of 0 prevents an analysis based on repeated χ^2 tests, but χ^2 can be run on the overall distributions of actions across the two groups within each project. Comparing the expected and observed distribution of actions within the Experimental group during the Learning project, there exists a statistically significant difference in the distributions ($p < 0.0001$, $\chi^2 = 60.307$, 9 degrees of freedom) (T.2L). The most significant differences come in a significantly fewer observed incidences of model construction than expected (−170) and notable spikes in the incidence of Revising Evidence (+53) and Revising Model (+154). As described before, there is not sufficient evidence to conclude differences along any one of these dimensions, but **there does exist sufficient evidence to state that there exists an overall difference of some kind between the Control and Experimental groups on the Learning project** derived from these individual differences. These observed differences may be strong parts of this overall difference.

Comparing the expected and observed distribution of actions within the Experimental group during the Transfer project, there also exists a statistically significant difference in the distributions ($p < 0.0001$, $\chi^2 = 188.084$, 9 degrees of freedom) (T.2T). The most significant differences here are a notable spike in the incidence of both Revising Model (+91) and Taking Notes (+26) relative to the values predicted by the Control group, as well as a notable drop in Using Simulation (−40). As with the results during the Learning project, there are not sufficient data to conclude any one of these differences is significant, but **there does exist sufficient evidence to state that there exists an overall difference of some kind between the Control and Experimental groups on the Transfer project** derived from these individual differences.

This analysis concludes the comparative analysis of the inquiry-driven modeling process between the Control and Experimental groups. Based on this analysis, there are sufficient data to conclude that based on interaction with MILA–T, there exists a difference in behavior between teams in the Control and Experimental groups. However, there do not exist sufficient data to conclude that the performance of teams in the Experimental group is *superior* to performance of teams in the Control group. Thus, **Hypothesis #3 is not accepted**. There are not sufficient data to conclude that engagement with MILA–T during the inquiry-driven modeling progress led to an improvement in the execution of inquiry-driven modeling itself.

Qualitative Analysis of the Tutors' Role

Despite the lack of data to conclude that engagement with MILA–T leads to improvement in the inquiry-driven modeling process, there is nonetheless sufficient information to support further analysis of the precise way in which the tutoring system was used during engagement with MILA. The above analysis asserts that a difference based on engagement with MILA–T does actually exist; it simply stops short of calling that difference an advantage. Moreover, the analysis of hypothesis #4, below, conclusively shows that teams in the Experimental condition produce better explanations than teams in the Control condition, suggesting that the process of model construction in the Experimental condition was superior even if that superiority did not come out in these specific metrics.

Examination of the Markov chain of the Experimental group during the Learning experiment including the Consulting Tutor action (shown previously in Markov Chain 1) yields a number of observations (T.3). First, interaction with the tutoring system is deeply embedded in interaction with MILA. 24.64% of all interactions with MILA by the Experimental group during the Learning project were interactions with the tutoring system. In total, teams in the Experimental group registered 1837 Consulting Tutor actions during the four days of the Learning project, an average of 37 interactions per team.

Consulting Tutors was a common action following every other action within MILA. It was the most common action to follow Dismissing Model and Writing Problem Definition, the second-most common action that teams took immediately after starting their projects (after Writing Problem Definition), and the second-most common action to take place after the Taking Notes action (after continued note-taking). Consulting Tutors is also among the most common actions to follow Constructing Evidence (second-most common), Constructing Model (third-most common), Proposing Hypothesis (second-most common), Reconsidering Model (second-most common), Revising Evidence (third-most common), Revising Model (third-most common), and Using Simulation (second-most common). Regardless of what teams did during interaction with MILA, consultation with one of the tutors often followed immediately thereafter.

The most common action to follow Consulting Tutor is a repeat of the Consulting Tutor action, suggesting multiple consecutive instances of requesting feedback from the Critic or asking the Guide a question. After repeated consultation with a tutor, however, teams most often moved toward model construction (19.9% of the remaining transitions), followed by evidence construction (15.4% of the remaining transitions). These transitions reflect some of the most common feedback that teams received, that their models' mechanisms were overly simplistic or their evidential justifications were overly reliant on weaker forms of evidence. As referenced earlier, the desired transition after receiving this feedback would be to gather additional information or evidence to add to the model, but after consultation with the tutors teams only transitioned to note-taking 4.1% of the remaining transitions and using a simulation 7.1% of the remaining transitions. As referenced previously, however, many information-gathering tasks were performed outside MILA, and thus this analysis does not infer that the lack of an explicit data-gathering task between the feedback and the continued model or evidence construction are instances of the error described previously.

Tutor-Specific Usage Patterns

The Markov chain featuring Consulting Tutors as an action can be further decomposed into a Markov chain reflecting interaction with each specific tutor. This Markov chain is difficult to visualize given the presence of sixteen different tasks (after splitting Consulting Tutor into the four specific tutors and the Tutor Feedback Received action), but the observations derived from this chain can be summarized. First, the chart in Figure 26 below provides the prevalence of interaction with each of the four tutors.

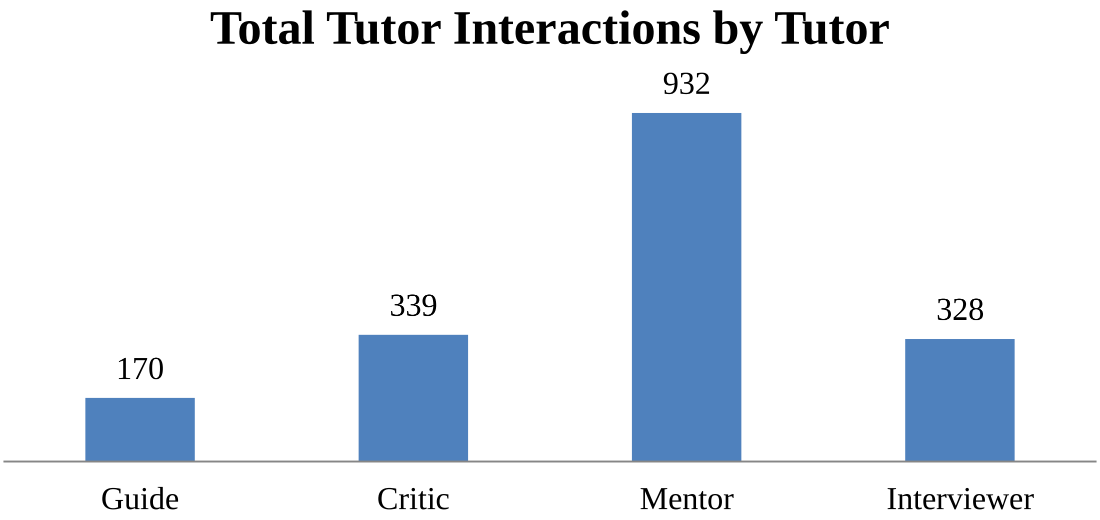


Figure 26: Distribution of tutor usage. The Mentor accounted for over half (52.7%) of all interactions with the tutoring system.

Interactions with the Mentor comprised roughly half of all interactions that teams had with the tutoring system. The Interviewer and Critic followed with approximately 20% of the interactions each, followed by the Guide with roughly 10%. This distribution of interaction with the tutoring system is somewhat expected; the Mentor initiates interaction with the teams, and therefore spawns more interaction than the Guide and Critic which rely on the team themselves to initiate interaction. The Interviewer also initiates interaction, but only based on a specific and much narrower set of prompts. While the Critic relies on teams to initiate interaction, it also provides more tangible immediate benefit as it responds directly to improvements made to the teams' models. In this way, the Critic is able to serve as an ongoing barometer of teams' success, thus spawning more interaction as teams check on their progress.

Each of the tutors shows distinct patterns of interaction. Interaction with the Mentor, for example, is most often followed by construction (15.45%) or revision (10.30%) of the team's models. Interaction with the Critic, on the other hand, is most often followed by addressing weaknesses in the model's evidential justifications (15.63% evidence construction, 2.95% evidence revision). The Critic is frequently followed by repeated interaction with the Critic (30.38%), reflecting that it gives different feedback when consulted repeatedly. Interaction with the Guide, too, is often followed by repeated interaction with the Guide (36.47%), suggesting teams navigating the question tree to get answers to different questions or perusing without a particular question in mind.

A significant feedback cycle also exists amongst the different tutors. Each tutor, at times, gives teams feedback on consulting with the other tutors. The Guide, for instance, points to the Critic as an example of model critique and the Mentor as a check for ongoing progress. The Mentor specifically comments on whether teams are consulting with the Guide and Critic, and the Interviewer asks teams to reflect on the usefulness and role of the tutors. As a result, a significant portion of interactions with the tutors are followed by interactions with other tutors: 22.12% of Critic interactions are followed by an interaction with one of the other tutors; this figure is 29.41% for the Guide, 15.65% for the Interviewer, and 20.39% for the Mentor. Similarly, 24.91% of the Critic's actions follow from one of the other tutors, along with 17.71% of the Guide's actions, 12.67% of the Interviewer's actions, and 32.27% of the Mentor's actions.

Although no firm statistical conclusions can be drawn based on patterns of interaction with the tutors due to the lack of a Control group, these observations regarding the pattern of interaction amongst the tutors are telling. First, the frequency of cycles within the tutoring system (the original Markov chain showed 56.34% of consultations with a tutor were followed immediately by another consultation with a tutor) suggests teams enter a general help-seeking phase during which they seek help from either multiple tutoring agents or the same agent multiple times. This suggests that the feedback teams are often seeking is beyond the level of a particular confusing part of their model, but rather reflects broader hesitation about what to do next. Second, the transitions out of the tutoring system suggest that teams attempt to immediately apply the feedback received. The majority of the feedback that the Critic has available to provide, for example, is for teams to improve the strength of their evidential justifications, and indeed, after receiving feedback from the Critic, that is exactly what teams most often do. The Mentor, on the other hand, comments more frequently on the specific structure of the team's model and their modeling process, and thus its feedback is followed more often by model construction and revision. As referenced above, the main reason these observations are interesting, however, is that the following analysis demonstrates that teams with MILA–T generate better explanations during both the Learning and the Transfer project; even if analysis of the process itself does not yield any clear differences, the output of the process is nonetheless superior for teams in the Experimental group based on some effect of the interactions documented here.

Hypothesis #4: Models and Explanations

The fourth hypothesis of this work stated that as a result of engagement with a metacognitive tutoring system during interaction with MILA, teams would generate superior models compared to teams that did not have the benefit of a tutoring system. This broad hypothesis broke down into two smaller hypotheses, dubbed the Learning hypothesis and the Transfer hypothesis. First, for the Learning hypothesis, teams who constructed models while interacting with MILA–T would construct superior models compared to teams who constructed their models without MILA–T. Second, for the Transfer hypothesis, teams who previously had constructed models while interacting with MILA–T would continue to construct superior models compared to teams who previously had constructed models without MILA–T, even after MILA–T was disabled. In other words, the gains seen through interaction with and feedback from MILA–T were internalized as learning gains that transferred to a new modeling task where such feedback was no longer available.

A number of analyses were conducted to evaluate the validity of these hypotheses. First, final models were analyzed as delivered by teams for their strength in terms of model complexity and evidential justifications, as described in Chapter 3. Second, the individual evidence that teams wrote in defense of their hypotheses and models was coded for accuracy along a number of different dimensions. This led to analysis of the overall quality of teams' written evidence, both across the aggregate total of all teams and across each individual team. Finally, these coded data were then used to reinterpret the original models that teams delivered according to the original metrics.

The process presented here is organized first according to the analysis, and then according to the specific hypothesis. Thus, first this section will discuss the initial analysis of models delivered by teams in both the Learning and Transfer projects. Then, this section will discuss the results of evidence coding in both the Learning and Transfer projects. Finally, this section will discuss the reinterpretation of the original models that teams delivered in both the Learning and Transfer projects.

Data Structure

Similar to Chapter 4, several tables will be presented throughout this chapter that summarize the numeric results from the Control and Experimental groups. For the sake of completeness, these tables present the mean, standard deviation, and number of teams within every significant group, from the Control and Experimental groups as a whole to narrower distinctions like Teacher A's On-Level Control group teams. Because of the quantity of data presented in these tables, they may be rather complex. In order to help understand these tables, the general structure of these data is presented below.

Table 47: The generic structure of the tables presented in this chapter. Each cell of these tables will provide three numbers: the mean on top, the standard deviation in parentheses in the middle, and the number of subjects at the bottom.

Generic Team Data Table Structure

<table>
<tr><td colspan="2">Teacher A</td><td colspan="2">Teacher B</td><td colspan="2">Teacher A</td><td colspan="2">Teacher B</td></tr>
<tr><td colspan="2">Teacher A's Control group teams</td><td colspan="2">Teacher B's Control group teams</td><td colspan="2">Teacher A's Experimental group teams</td><td colspan="2">Teacher B's Experimental group teams</td></tr>
<tr><td colspan="4">Control</td><td colspan="4">Experimental</td></tr>
<tr><td colspan="4">All Control group teams</td><td colspan="4">All Experimental group teams</td></tr>
<tr><td colspan="2">On-Level</td><td colspan="2">Gifted</td><td colspan="2">On-Level</td><td colspan="2">Gifted</td></tr>
<tr><td colspan="2">On-Level Control group teams</td><td colspan="2">Gifted Control group teams</td><td colspan="2">On-Level Experimental group teams</td><td colspan="2">Gifted Experimental group teams</td></tr>
<tr><td>A</td><td>B</td><td>A</td><td>B</td><td>A</td><td>B</td><td>A</td><td>B</td></tr>
<tr><td>Teacher A's On-Level Control group teams</td><td>Teacher B's On-Level Control group teams</td><td>Teacher A's Gifted Control group teams</td><td>Teacher B's Gifted Control group teams</td><td>Teacher A's On-Level Exper-imental group teams</td><td>Teacher B's On-Level Exper-imental group teams</td><td>Teacher A's Gifted Exper-imental group teams</td><td>Teacher B's Gifted Exper-imental group teams</td></tr>
</table>

Each cell in the tables that follow provides the subscores for the group of teams indicated in Table 47. Within each cell are three numbers: the number at the top is the **mean** for that group of teams; the number in parentheses in the middle is the **standard deviation** for that group of teams; and the number at the bottom, followed by "*n* =", is the **number of teams** in that group of teams. Teacher B has no Gifted classes, and as such the 'Teacher B's Gifted' team cells are empty in subsequent tables. These tables are similar to those used in Chapter 4; however, given that Special Needs students were mingled with On-Level students in On-Level classes, there is no 'special needs' category for teams. The table is structured to allow as many comparisons as possible under the general structure of the experiment. We can, for instance, directly compare the Control and Experimental teams as a whole by comparing the 'All Control group teams' and 'All Experimental group teams' cells, or we may compare only Teacher A's teams by comparing the 'Teacher A's Control group teams' and 'Teacher A's Experimental group teams'. In reading the tables that follow, please consult Table 47 to ensure understanding of the structure of the data presented.

Initial Model Analysis

Initial analysis of the models that teams submitted examined the models along five separate metrics. The first metric, Model Complexity, summarized the complexity and interconnectedness of teams' models. The Model Complexity metric summed the number of components and connections in teams' models without weighting them to capture the size of models in both number of interacting variables and components and number of actual interactions between those variables and components. The other four metrics all built from different specific elements of the evidential justifications teams provided within their models: Total Model Strength, Average Model Strength, Average Evidence Strength, and Total Evidence. Total Model Strength compiled the evidence across teams' entire models according to the weighting scheme described earlier. For example, a model supplying two pieces of Expert Information (3 points each) and two Direct Observation (1 point each) would receive a score of '8'. Average Model Strength averaged this total strength across the number of edges in the model; for example, if there were two edges in the model, the previous model would have an Average Model Strength of 4. This metric controls for instances where models had greater total strength because of a greater number of claims made rather than greater evidence for individual claims. Average Evidence Strength took this control structure a step further and examined not only the average across the total number of claims made, but also across the total number of individual pieces of evidence supplied in justification of those claims. In the previous example, the Average Evidence Strength would be 2 (8 points over 4 pieces of evidence). This metric controls for instances wherein teams provided a greater number of weak pieces of evidence to arrive at the same overall strength (for example, lots of Logical

Explanations), rather than a smaller number of strong pieces of evidence (for example, a few pieces of Expert Information). Finally, the Total Evidence metric serves as a control over all of these to see if any observed differences in the above variables could be attributed solely to a change in the raw number of pieces of evidence supplied by teams rather than variations in the actual strength of that evidence. In the above example, 4 pieces of evidence were supplied, so the Total Evidence score would be 4.

The granularity of these metrics aims to draw a tight conclusion regarding the mechanism by which any differences that are observed arise. If, for example, teams in the Experimental group demonstrated a significantly higher overall strength of evidence in their models, that greater strength could be derived from multiple different changes, such as larger, more complex models (providing more claims to justify with evidence and, thus, more evidence); more individual pieces of evidence without a difference in the strength of that evidence; or a preference for actual stronger pieces of evidence.

Learning Project

At the conclusion of the Learning project, each team was asked to supply one model that represented their best explanation for what caused the fish kill in Lake Clara Meer; teams were told that they would be turning in a single model at the conclusion of their inquiry two days prior to the conclusion of the project. Initially, a multivariate analysis was conducted on these models to discern differences between teams' models. Three independent variables were considered, Condition, Teacher, and Class Level (gifted or on-level), and five output variables were considered as described above (Model Complexity, Total Model Strength, Average Model Strength, Average Evidence Strength, and Total Evidence). The objective of this analysis was to identify which independent variables predicted differences in teams' models across these five metrics.

Initial multivariate analysis of variance revealed a statistically significant effect of Condition (F = 3.04, $p < 0.05$), Teacher (F = 3.28, $p < 0.01$), and Class Level (F = 7.54, $p \approx 0.0$) (T.4L). The results of this initial analysis of variance are shown below in Table 48.

Table 48: Multivariate ANOVA results on the differences in teams' models across five metrics based on different independent variables. The results indicate statistically significant effects of all three independent variables: Condition (F = 3.04, *p* < 0.05), Teacher (F = 3.28, *p* < 0.01), and Class Level (F = 7.54, *p* ≈ 0.0)

```
                     Df  Pillai approx F num Df den Df    Pr(>F)
Condition             1 0.17019   3.0355      5     74  0.015116
Teacher               1 0.18142   3.2801      5     74  0.009916
ClassLevel            1 0.33756   7.5416      5     74 9.087e-06
Condition:Teacher     1 0.03129   0.4781      5     74  0.791516
Condition:ClassLevel  1 0.05205   0.8127      5     74  0.544410
Residuals            78
```

Based on the results of this analysis, a follow-up univariate analysis of variance was conducted on each of the five metrics. The raw results of this univariate analysis are shown below in Table 49.

A number of conclusions can be drawn based on these univariate analyses (T.4L.A-E). First and foremost with regard to the hypotheses and Experimental condition, a statistically significant effect was seen based on the Condition along three of the five metrics: Total Strength (F = 6.71, $p < 0.05$), Average Strength (F = 4.23, $p < 0.05$), and Average Evidence Strength (F = 4.70, $p < 0.05$). No statistically significant effect of the Condition is seen on Model Complexity (F = 0.89, $p = 0.35$) or Total Evidence (F = 1.96, $p = 0.17$). These results show that **teams in the Experimental group delivered models that provided stronger evidence for their claims than teams in the Control group**. The difference in Total Strength demonstrated this overall difference first, as teams in the Experimental group received an average Total Evidence Strength score of 17.24 ($\sigma = 13.24$), compared to 11.03 ($\sigma = 7.67$) for teams in the Control group. In other words, models generated by teams in the Experimental group were an average of 6.21 points, or 56.3%, stronger than models generated by teams in the Control group.

Table 49: Univariate ANOVA results each of the five metrics describing the differences in teams' models across the different independent variables.

```
Response P1.ModelComplexity :
                     Df Sum Sq Mean Sq F value    Pr(>F)
Condition             1   6.95   6.947  0.8914    0.3480
Teacher               1  19.64  19.639  2.5200    0.1165
ClassLevel            1 285.34 285.336 36.6130 4.709e-08
Condition:Teacher     1   6.11   6.112  0.7843    0.3786
Condition:ClassLevel  1   1.37   1.374  0.1763    0.6757
Residuals            78 607.88   7.793

 Response P1.TotalStrength :
                     Df Sum Sq Mean Sq F value   Pr(>F)
Condition             1  780.6  780.61  6.7111 0.011431
Teacher               1  356.6  356.55  3.0654 0.083907
ClassLevel            1 1053.9 1053.88  9.0605 0.003517
Condition:Teacher     1   47.6   47.58  0.4090 0.524344
Condition:ClassLevel  1    1.4    1.42  0.0122 0.912230
Residuals            78 9072.7  116.32

 Response P1.AverageStrength :
                     Df Sum Sq Mean Sq F value   Pr(>F)
Condition             1  21.57  21.568  4.2308 0.043040
Teacher               1  49.53  49.530  9.7157 0.002558
ClassLevel            1   2.44   2.438  0.4783 0.491264
Condition:Teacher     1   1.56   1.555  0.3050 0.582314
Condition:ClassLevel  1   0.67   0.670  0.1315 0.717846
Residuals            78 397.64   5.098

 Response P1.TotalEvidence :
                     Df  Sum Sq Mean Sq F value   Pr(>F)
Condition             1   56.56  56.561  1.9585 0.165638
Teacher               1  165.50 165.498  5.7306 0.019074
ClassLevel            1  300.00 300.004 10.3880 0.001852
Condition:Teacher     1   15.36  15.359  0.5318 0.468022
Condition:ClassLevel  1    2.36   2.361  0.0818 0.775682
Residuals            78 2252.63  28.880

 Response P1.AverageEvidenceStrength :
                     Df  Sum Sq Mean Sq F value  Pr(>F)
Condition             1  1.8648 1.86484  4.7012 0.03319
Teacher               1  0.3454 0.34539  0.8707 0.35364
ClassLevel            1  1.9683 1.96830  4.9620 0.02879
Condition:Teacher     1  0.0094 0.00939  0.0237 0.87810
Condition:ClassLevel  1  0.0438 0.04381  0.1104 0.74053
Residuals            78 30.9405 0.39667
```

This finding could, again, be explained in two different ways: either teams in the Experimental group used a greater number of equally-strong pieces of evidence to defend their models, or teams in the Experimental group used an equal number of pieces of evidence while making the individual pieces of evidence stronger. Based on the lack of difference in Total Evidence and the present difference in Average Evidence Strength between the Control and Experimental groups, the latter explanation is accurate. There are not data indicating that teams in the Experimental group simply used a greater number of pieces of evidence. Instead, teams in the Experimental group received an Average Evidence Strength score of 1.64 (σ = 0.69), while teams in the Control group received an Average Evidence Strength score of 1.34 (σ = 0.56). In other words, each individual piece of evidence that teams in the Experimental group supplied was an average of 0.30 points, or 22%, stronger than each individual piece of evidence that teams in the Control group supplied according to the weighting scheme given in Chapter 3.

The numeric results for each metric for each group are shown in the tables below.

Table 50: Difference in Model Complexity between teams in different Conditions, Teachers, Class Levels, and combinations thereof. The top number in each cell is the mean for that group; the number in parentheses is the standard deviation within that group; the bottom number is the number of samples within that subgroup. See Table 47 for further information about the groups of students summarized in each cell.

Difference in Model Complexity in the Learning Project

<table>
<tr><td colspan="2">Teacher A</td><td colspan="2">Teacher B</td><td colspan="2">Teacher A</td><td colspan="2">Teacher B</td></tr>
<tr><td colspan="2">10.25
(4.39)
n = 16</td><td colspan="2">8.44
(1.92)
n = 18</td><td colspan="2">10.08
(3.72)
n = 25</td><td colspan="2">9.68
(2.94)
n = 25</td></tr>
<tr><td colspan="4">Control</td><td colspan="4">Experimental</td></tr>
<tr><td colspan="4">9.29
(3.39)
n = 34</td><td colspan="4">9.88
(3.32)
n = 50</td></tr>
<tr><td colspan="2">On-Level</td><td colspan="2">Gifted</td><td colspan="2">On-Level</td><td colspan="2">Gifted</td></tr>
<tr><td colspan="2">8.29
(2.26)
n = 28</td><td colspan="2">14.00
(4.00)
n = 6</td><td colspan="2">9.17
(2.74)
n = 42</td><td colspan="2">13.62
(3.78)
n = 8</td></tr>
<tr><td>A</td><td>B</td><td>A</td><td>B</td><td>A</td><td>B</td><td>A</td><td>B</td></tr>
<tr><td>8.00
(2.87)
n = 10</td><td>8.44
(1.92)
n = 18</td><td>14.00
(4.00)
n = 6</td><td>--
--
n = 0</td><td>8.41
(2.29)
n = 17</td><td>9.68
(2.94)
n = 25</td><td>13.62
(3.78)
n = 8</td><td>--
--
n = 0</td></tr>
</table>

Table 51: Difference in Total Model Strength between teams in different Conditions, Teachers, Class Levels, and combinations thereof. The top number in each cell is the mean for that group; the number in parentheses is the standard deviation within that group; the bottom number is the number of samples within that subgroup. See Table 47 for further information about the groups of students summarized in each cell.

Difference in Total Model Strength in the Learning Project

Teacher A		Teacher B		Teacher A		Teacher B	
10.00 (7.18) *n* = 16		11.94 (8.16) *n* = 18		14.44 (13.71) *n* = 25		20.04 (12.40) *n* = 25	
Control				Experimental			
11.03 (7.67) *n* = 34				17.24 (13.24) *n* = 50			
On-Level		Gifted		On-Level		Gifted	
9.89 (7.54) *n* = 28		16.33 (6.28) *n* = 6		16.36 (13.13) *n* = 42		21.88 (13.76) *n* = 8	
A	B	A	B	A	B	A	B
6.20 (4.61) *n* = 10	11.94 (8.16) *n* = 18	16.33 (6.28) *n* = 6	-- -- *n* = 0	10.94 (12.59) *n* = 17	20.04 (12.4) *n* = 25	21.88 (13.76) *n* = 8	-- -- *n* = 0

Table 52: Difference in Average Model Strength between teams in different Conditions, Teachers, Class Levels, and combinations thereof. The top number in each cell is the mean for that group; the number in parentheses is the standard deviation within that group; the bottom number is the number of samples within that subgroup. See Table 47 for further information about the groups of students summarized in each cell.

Difference in Average Model Strength in the Learning Project

Teacher A		Teacher B		Teacher A		Teacher B	
1.96 (1.39) *n* = 16		3.15 (2.05) *n* = 18		2.73 (2.15) *n* = 25		4.51 (2.81) *n* = 25	
Control				Experimental			
2.59 (1.84) *n* = 34				3.62 (2.63) *n* = 50			
On-Level		Gifted		On-Level		Gifted	
2.61 (1.93) *n* = 28		2.47 (1.50) *n* = 6		3.75 (2.83) *n* = 42		2.92 (0.99) *n* = 8	
A	B	A	B	A	B	A	B
1.65 (1.30) *n* = 10	3.15 (2.05) *n* = 18	2.47 (1.50) *n* = 6	-- -- *n* = 0	2.65 (2.55) *n* = 17	4.51 (2.81) *n* = 25	2.92 (0.99) *n* = 8	-- -- *n* = 0

Table 53: Difference in Average Evidence Strength between teams in different Conditions, Teachers, Class Levels, and combinations thereof. The top number in each cell is the mean for that group; the number in parentheses is the standard deviation within that group; the bottom number is the number of samples within that subgroup. See Table 47 for further information about the groups of students summarized in each cell.

Difference in Average Evidence Strength in the Learning Project

Teacher A		Teacher B		Teacher A		Teacher B	
1.26 (0.56) *n* = 16		1.40 (0.56) *n* = 18		1.58 (0.81) *n* = 25		1.70 (0.54) *n* = 25	
Control				Experimental			
1.34 (0.56) *n* = 34				1.64 (0.69) *n* = 50			
On-Level		Gifted		On-Level		Gifted	
1.30 (0.60) *n* = 28		1.50 (0.24) *n* = 6		1.59 (0.72) *n* = 42		1.93 (0.34) *n* = 8	
A	B	A	B	A	B	A	B
1.12 (0.66) *n* = 10	1.40 (0.56) *n* = 18	1.50 (0.24) *n* = 6	-- -- *n* = 0	1.41 (0.92) *n* = 17	1.70 (0.54) *n* = 25	1.93 (0.34) *n* = 8	-- -- *n* = 0

Table 54: Difference in Total Evidence between teams in different Conditions, Teachers, Class Levels, and combinations thereof. The top number in each cell is the mean for that group; the number in parentheses is the standard deviation within that group; the bottom number is the number of samples within that subgroup. See Table 47 for further information about the groups of students summarized in each cell.

Difference in Total Evidence in the Learning Project

Teacher A		Teacher B		Teacher A		Teacher B	
6.75 (4.52) *n* = 16		8.33 (4.52) *n* = 18		7.44 (6.04) *n* = 25		11.08 (6.51) *n* = 25	
Control				Experimental			
7.59 (4.53) *n* = 34				9.26 (6.48) *n* = 50			
On-Level		Gifted		On-Level		Gifted	
6.93 (4.50) *n* = 28		10.67 (3.50) *n* = 6		8.93 (6.64) *n* = 42		11.00 (5.66) *n* = 8	
A	B	A	B	A	B	A	B
4.40 (3.31) *n* = 10	8.33 (4.52) *n* = 18	10.67 (3.50) *n* = 6	-- -- *n* = 0	5.76 (5.61) *n* = 17	11.08 (6.51) *n* = 25	11.00 (5.66) *n* = 8	-- -- *n* = 0

To summarize the findings above, statistically significant differences based on the Condition were observed along three metrics: Total Model Strength, Average Model Strength, and Average Evidence Strength. Based on these differences, this analysis shows that **teams in the Experimental group produced better-justified models on the Learning project than teams in the Control group**, as measured by the metrics noted above.

In addition, a number of other interactions are worth noting based on this analysis. First, there existed a statistically significant interaction between four of the five metrics and Class Level: Model Complexity (F = 36.6, $p \approx 0.0$), Total Model Strength (F = 9.06, $p < 0.01$), Total Evidence (F = 10.39, $p < 0.01$), and Average Evidence Strength (F = 4.96, $p < 0.05$). In each, teams in the Gifted classes outperformed teams in the On-Level classes. However, the absence of an effect in Average Evidence Strength suggests that these effects were largely a product of producing bigger models with more evidence rather than producing stronger evidence. In fact, although the difference is not statistically significant, teams in the On-Level classes had a higher Average Evidence Strength (3.30, $\sigma = 2.56$) than teams in Gifted classes (2.73, $\sigma = 1.20$). Secondly, there existed a statistically significant interaction between Teacher and two metrics: Average Model Strength (F = 9.72, $p < 0.01$) and Total Evidence (F = 5.73, $p < 0.05$). In both metrics, teams in Teacher B's classes outperformed teams in Teacher A's classes.

Transfer Project

At the conclusion of the Transfer project, each team was asked to supply one model that represented their best explanation of what has caused Atlanta's recent run of record-high temperatures. Teams were informed at the beginning of the project that they would be delivering such an explanation at the conclusion of class. Initially, a multivariate analysis was conducted on these models to discern differences between teams' models. Three independent variables were considered, Condition, Teacher, and

Class Level (gifted or on-level), and five output variables were considered as described above (Model Complexity, Total Model Strength, Average Model Strength, Average Evidence Strength, and Total Evidence). The objective of this analysis was to identify which independent variables predicted differences in teams' models across these five metrics.

Initial multivariate analysis of variance revealed a statistically significant effect of Condition (F = 3.21, $p < 0.05$) and Teacher (F = 4.02, $p < 0.01$), but not of Class Level (F = 1,11, $p = 0.36$) (T.4T). The raw results of the initial multivariate analysis of variance are shown below in Table 55.

Table 55: Multivariate ANOVA results on the differences in teams' Transfer project models across five metrics based on different independent variables. The results indicate statistically significant effects of all three independent variables: Condition (F = 3.21, *p* < 0.05) and Teacher (F = 4.02, *p* < 0.01).

```
                     Df   Pillai approx F num Df den Df   Pr(>F)
Condition             1 0.184474   3.2121      5     71 0.011360
Teacher               1 0.220778   4.0233      5     71 0.002854
ClassLevel            1 0.072455   1.1092      5     71 0.363306
Condition:Teacher     1 0.082125   1.2705      5     71 0.286239
Condition:ClassLevel  1 0.024373   0.3547      5     71 0.877562
Residuals            75
```

Based on the results of this analysis, a follow-up univariate analysis of variance was conducted on each of the five metrics. The raw results of this univariate analysis are shown below in Table 56.

With regard to the hypotheses and Experimental condition, a statistically significant effect was seen based on the Condition along three of the five metrics: Model Complexity (F = 4.95, $p < 0.05$), Total Model Strength (F = 4.61, $p < 0.05$), and Average Evidence Strength (F = 9.48, $p < 0.01$) (T.4T.A-E). No statistically significant effect of the Condition is seen on Average Model Strength (F = 0.87, $p = 0.35$) or Total Evidence (F = 0.66, $p = 0.42$). Unlike in the Learning project, the Transfer project did see a difference in Model Complexity between the Control and Experimental groups: teams in the Experimental group received an average Model Complexity score of 10.62 ($\sigma = 3.70$), while teams in the Control group received an average score of 8.91 ($\sigma = 4.06$), a 1.71

point advantage for teams in the Experimental group. For comparison, teams in the Experimental group received an average Model Complexity score of 9.88 (σ = 3.32) in the Learning project, while teams in the Control group received an average score of 9.29 (σ = 3.39).

Table 56: Univariate ANOVA results each of the five metrics describing the differences in teams' Transfer project models across the different independent variables.

```
Response P2.ModelComplexity :
                     Df Sum Sq Mean Sq F value    Pr(>F)
Condition             1  57.37  57.368  4.9485   0.02912
Teacher               1 209.35 209.353 18.0585 6.088e-05
ClassLevel            1  58.96  58.962  5.0860   0.02704
Condition:Teacher     1  31.54  31.544  2.7210   0.10322
Condition:ClassLevel  1   4.51   4.508  0.3888   0.53481
Residuals            75 869.48  11.593

 Response P2.TotalStrength :
                     Df Sum Sq Mean Sq F value Pr(>F)
Condition             1  320.4  320.41  4.6109 0.0350
Teacher               1  127.8  127.82  1.8394 0.1791
ClassLevel            1   87.6   87.60  1.2606 0.2651
Condition:Teacher     1   12.4   12.41  0.1786 0.6738
Condition:ClassLevel  1   85.2   85.25  1.2268 0.2716
Residuals            75 5211.8   69.49

 Response P2.AverageStrength :
                     Df Sum Sq Mean Sq F value Pr(>F)
Condition             1   3.95  3.9547  0.8676 0.3546
Teacher               1   6.17  6.1736  1.3543 0.2482
ClassLevel            1   0.09  0.0884  0.0194 0.8896
Condition:Teacher     1   2.22  2.2156  0.4860 0.4879
Condition:ClassLevel  1   2.71  2.7142  0.5954 0.4427
Residuals            75 341.88  4.5584

 Response P2.TotalEvidence :
                     Df  Sum Sq Mean Sq F value Pr(>F)
Condition             1   11.80 11.8025  0.6630 0.4181
Teacher               1   29.87 29.8666  1.6778 0.1992
ClassLevel            1   18.80 18.8016  1.0562 0.3074
Condition:Teacher     1    0.25  0.2488  0.0140 0.9062
Condition:ClassLevel  1    9.08  9.0850  0.5104 0.4772
Residuals            75 1335.08 17.8011

 Response P2.AverageEvidenceStrength :
                     Df Sum Sq Mean Sq F value   Pr(>F)
Condition             1  5.006  5.0062  9.4769 0.002906
Teacher               1  0.205  0.2050  0.3881 0.535169
ClassLevel            1  0.035  0.0349  0.0661 0.797852
Condition:Teacher     1  0.216  0.2156  0.4081 0.524897
Condition:ClassLevel  1  0.182  0.1817  0.3440 0.559287
Residuals            75 39.619  0.5282
```

As in the Learning project, a difference in Total Model Strength between teams in the Control and Experimental group could be explained in one of two ways: either the teams with the higher strength had larger models, or they could have models of the same size with stronger evidence. Unlike in the Learning project, however, here the first explanation holds true. Teams in the Experimental group received an average Total Model Strength score of of 14.38 (σ = 9.09), while teams in the Control group received an average score of 10.35 (σ = 7.24), a 4.03 point difference in favor of teams in the Experimental group. Combined with the increased score on the Model Complexity metric, this result establishes that **teams in the Experimental group were able to build larger, more complex models of the phenomenon without sacrificing the evidential strength underlying their claims**. Put more simply, in the Learning project, teams in the Experimental group did a better job of justifying their claims; **in the Transfer project, teams in the Experimental group justified more claims than the Control group equally well**. Table 57 through Table 61 show the full results of this analysis.

In order to justify more claims, teams in the Experimental group would have to do one of two things: either they would have to supply more bits of evidence overall, or the bits of evidence they supplied would have to be stronger. Given that the models from teams in the Experimental group were more complex, it would stand to reason that they justified those models with a greater number of bits of evidence. However, the analysis revealed no significant difference in the Total Evidence metric between teams in the Control and Experimental groups; teams in the Experimental group supplied on average 7.36 (σ = 4.20) pieces of evidence per model, while teams in the Control group supplied an average of 6.59 (σ = 4.20) pieces, a minor difference. Rather, the greater Total Model Strength and matching Average Model Strength are owed to the Experimental group providing stronger individual pieces of evidence; the mean Average Evidence Strength score for teams in the Experimental group was 1.96 (σ = 0.70), while for teams in the Control group it was 1.46 (σ = 0.74).

Table 57: Difference in Model Complexity between teams in different Conditions, Teachers, Class Levels, and combinations thereof. The top number in each cell is the mean for that group; the number in parentheses is the standard deviation within that group; the bottom number is the number of samples within that subgroup. See Table 47 for further information about the groups of students summarized in each cell.

Difference in Model Complexity in the Transfer Project

<table>
<tr><td colspan="2">Teacher A</td><td colspan="2">Teacher B</td><td colspan="2">Teacher A</td><td colspan="2">Teacher B</td></tr>
<tr><td colspan="2">11.44
(3.85)
n = 16</td><td colspan="2">6.67
(2.74)
n = 18</td><td colspan="2">11.60
(3.58)
n = 25</td><td colspan="2">9.50
(3.60)
n = 22</td></tr>
<tr><td colspan="4">Control</td><td colspan="4">Experimental</td></tr>
<tr><td colspan="4">8.91
(4.06)
n = 34</td><td colspan="4">10.62
(3.70)
n = 47</td></tr>
<tr><td colspan="2">On-Level</td><td colspan="2">Gifted</td><td colspan="2">On-Level</td><td colspan="2">Gifted</td></tr>
<tr><td colspan="2">7.93
(3.48)
n = 28</td><td colspan="2">13.50
(3.51)
n = 6</td><td colspan="2">10.15
(3.83)
n = 39</td><td colspan="2">12.88
(1.89)
n = 8</td></tr>
<tr><td>A</td><td>B</td><td>A</td><td>B</td><td>A</td><td>B</td><td>A</td><td>B</td></tr>
<tr><td>10.20
(3.65)
n = 10</td><td>6.67
(2.74)
n = 18</td><td>13.50
(3.51)
n = 6</td><td>--
--
n = 0</td><td>11.00
(4.06)
n = 17</td><td>9.50
(3.60)
n = 22</td><td>12.88
(1.89)
n = 8</td><td>--
--
n = 0</td></tr>
</table>

Table 58: Difference in Total Model Strength between teams in different Conditions, Teachers, Class Levels, and combinations thereof. The top number in each cell is the mean for that group; the number in parentheses is the standard deviation within that group; the bottom number is the number of samples within that subgroup. See Table 47 for further information about the groups of students summarized in each cell.

Difference in Total Model Strength in the Transfer Project

<table>
<tr><td colspan="2">Teacher A</td><td colspan="2">Teacher B</td><td colspan="2">Teacher A</td><td colspan="2">Teacher B</td></tr>
<tr><td colspan="2">11.25
(7.37)
n = 16</td><td colspan="2">9.56
(7.23)
n = 18</td><td colspan="2">15.84
(10.91)
n = 25</td><td colspan="2">12.73
(6.28)
n = 22</td></tr>
<tr><td colspan="4">Control</td><td colspan="4">Experimental</td></tr>
<tr><td colspan="4">10.35
(7.24)
n = 34</td><td colspan="4">14.38
(9.09)
n = 47</td></tr>
<tr><td colspan="2">On-Level</td><td colspan="2">Gifted</td><td colspan="2">On-Level</td><td colspan="2">Gifted</td></tr>
<tr><td colspan="2">9.25
(6.89)
n = 28</td><td colspan="2">15.50
(7.12)
n = 6</td><td colspan="2">14.00
(9.42)
n = 39</td><td colspan="2">16.25
(7.50)
n = 8</td></tr>
<tr><td>A</td><td>B</td><td>A</td><td>B</td><td>A</td><td>B</td><td>A</td><td>B</td></tr>
<tr><td>8.70
(6.57)
n = 10</td><td>9.56
(7.23)
n = 18</td><td>15.50
(7.12)
n = 6</td><td>--
--
n = 0</td><td>15.65
(12.40)
n = 17</td><td>12.73
(6.28)
n = 22</td><td>16.25
(7.50)
n = 8</td><td>--
--
n = 0</td></tr>
</table>

Table 59: Difference in Average Model Strength between teams in different Conditions, Teachers, Class Levels, and combinations thereof. The top number in each cell is the mean for that group; the number in parentheses is the standard deviation within that group; the bottom number is the number of samples within that subgroup. See Table 47 for further information about the groups of students summarized in each cell.

Difference in Average Model Strength in the Transfer Project

Teacher A		Teacher B		Teacher A		Teacher B	
2.25 (1.60) *n* = 16		3.19 (2.15) *n* = 18		3.07 (2.40) *n* = 25		3.34 (2.06) *n* = 22	
Control				Experimental			
2.75 (1.95) *n* = 34				3.20 (2.23) *n* = 47			
On-Level		Gifted		On-Level		Gifted	
2.78 (2.06) *n* = 28		2.61 (1.41) *n* = 6		3.30 (2.36) *n* = 39		2.71 (1.45) *n* = 8	
A	B	A	B	A	B	A	B
2.03 (1.75) *n* = 10	3.19 (2.15) *n* = 18	2.61 (1.41) *n* = 6	-- -- *n* = 0	3.24 (2.76) *n* = 17	3.34 (2.06) *n* = 22	2.71 (1.45) *n* = 8	-- -- *n* = 0

Table 60: Difference in Average Evidence Strength between teams in different Conditions, Teachers, Class Levels, and combinations thereof. The top number in each cell is the mean for that group; the number in parentheses is the standard deviation within that group; the bottom number is the number of samples within that subgroup. See Table 47 for further information about the groups of students summarized in each cell.

Difference in Average Evidence Strength in the Transfer Project

Teacher A		Teacher B		Teacher A		Teacher B	
1.58 (0.55) *n* = 16		1.35 (0.87) *n* = 18		1.97 (0.58) *n* = 25		1.96 (0.83) *n* = 22	
Control				Experimental			
1.46 (0.74) *n* = 34				1.96 (0.70) *n* = 47			
On-Level		Gifted		On-Level		Gifted	
1.40 (0.80) *n* = 28		1.72 (0.15) *n* = 6		1.97 (0.74) *n* = 39		1.93 (0.47) *n* = 8	
A	B	A	B	A	B	A	B
1.49 (0.68) *n* = 10	1.35 (0.87) *n* = 18	1.72 (0.15) *n* = 6	-- -- *n* = 0	1.99 (0.64) *n* = 17	1.96 (0.83) *n* = 22	1.93 (0.47) *n* = 8	-- -- *n* = 0

Table 61: Difference in Total Evidence between teams in different Conditions, Teachers, Class Levels, and combinations thereof. The top number in each cell is the mean for that group; the number in parentheses is the standard deviation within that group; the bottom number is the number of samples within that subgroup. See Table 47 for further information about the groups of students summarized in each cell.

Difference in Total Evidence in the Transfer Project

<table>
<tr><td colspan="2">Teacher A</td><td colspan="2">Teacher B</td><td colspan="2">Teacher A</td><td colspan="2">Teacher B</td></tr>
<tr><td colspan="2">7.19
(4.56)
n = 16</td><td colspan="2">6.06
(3.90)
n = 18</td><td colspan="2">7.96
(4.64)
n = 25</td><td colspan="2">6.68
(3.62)
n = 22</td></tr>
<tr><td colspan="4">Control</td><td colspan="4">Experimental</td></tr>
<tr><td colspan="4">6.59
(4.20)
n = 34</td><td colspan="4">7.36
(4.20)
n = 47</td></tr>
<tr><td colspan="2">On-Level</td><td colspan="2">Gifted</td><td colspan="2">On-Level</td><td colspan="2">Gifted</td></tr>
<tr><td colspan="2">6.11
(4.18)
n = 28</td><td colspan="2">8.83
(3.87)
n = 6</td><td colspan="2">7.15
(4.32)
n = 39</td><td colspan="2">8.38
(3.62)
n = 8</td></tr>
<tr><td>A</td><td>B</td><td>A</td><td>B</td><td>A</td><td>B</td><td>A</td><td>B</td></tr>
<tr><td>6.20
(4.85)
n = 10</td><td>6.06
(3.90)
n = 16</td><td>8.83
(3.87)
n = 6</td><td>--
--
n = 0</td><td>7.76
(5.14)
n = 17</td><td>6.68
(3.62)
n = 22</td><td>8.38
(3.62)
n = 8</td><td>--
--
n = 0</td></tr>
</table>

Summary

In the Learning project, teams with access to MILA–T provided models that were significantly higher in Total Model Strength, Average Model Strength, and Average Evidence Strength than teams that conducted created their models without MILA–T. Teams in the two groups produced models that were equally large and complex, and also provided the same raw number of pieces of evidence in support of those models. However, the individual pieces of evidence in defense of these models provided by teams using MILA–T were stronger than those provided by other teams. As a result, teams with MILA–T outperformed teams without MILA–T across all three metrics.

In the Transfer project, all teams operated without MILA–T. Teams that had previously had access to MILA–T during the Learning project, however, provided models that were significantly higher in Model Complexity, Total Model Strength, and Average Evidence Strength. Thus, teams that had previously used MILA–T were able to justify more claims in their explanations than teams that had not previously used MILA–T. Additionally, these teams did so with stronger evidence; teams that had previously used MILA–T were able to justify larger, more complex explanations with stronger pieces of evidence.

Coded Evidence Analysis

Based on the analysis of the models that teams submitted at the end of the Learning and Transfer project, improvement based on engagement with MILA–T is reasonably well-demonstrated. However, two questions remain after that analysis. First, and to verify the validity of that claim, was the evidence that teams supplied in defense of their explanations really better in the Experimental group than in the Control group? The scores associated with evidence strength in the prior analysis were based on teams' self-reported categorization of their evidence. It is possible that the teams could improve in their scores along these metrics simply by choosing stronger categories for their evidence rather than actually changing the evidence itself. Thus, it is important to examine the actual text of the evidence supplied by teams in addition to the categories.

Secondly, an additional conclusion is possible based on interpretation of the actual textual evidence that teams supplied. While the above conclusions and claims suggest a shift in the type of evidence that teams prefer, they do not address the deeper understanding of what evidence is, how it should be used, and what constitutes a strong defense beyond simply the right combination of categories. In addition to verifying that the evidence teams supplied for their explanations is itself sound to validate the previous conclusions, this analysis also aims to assess whether or not MILA–T helped teams develop a better understanding of the role of evidence and justifying claims in general.

Coding the evidence that teams wrote for their models consisted of three phases. First, a subset of the data was analyzed and coded in order to establish a coding scheme to use on the evidence as a whole. Then, three rounds of coding were run, each completed by the same coder three weeks apart. Intra-coder reliability was then established across these three coding sessions. Finally, the results of this exercise were analyzed as well to determine whether participation in the Experimental condition changed teams' performance in the quality, accuracy, and strength of evidence supplied for their explanations.

The teams in the study produced a total of 1301 individual pieces of evidence over the course of the two projects: 722 for the Learning project (261 Control, 461 Experimental) and 580 for the Transfer project (231 Control, 349 Experimental). These 1301 individual pieces of evidence were used to support 640 total claims or connections: 332 for the Learning project (125 Control, 207 Experimental) and 308 for the Transfer project (109 Control, 199 Experimental). Each piece of evidence was coded based on the text of the evidence, the category chosen by teams, and the connection or claim in the model that teams defended using that piece of evidence. Coding was randomized across all claims, although all pieces of evidence associated with a single connection were coded at the same time to identify instances of redundant justification.

Coding Scheme

Initially, 15% of the individual connections (96 connections total) were randomly selected and separated out. These 96 connections carried 199 individual pieces of evidence. Notes were taken on the acceptability of the evidence, and in the event of unacceptability, the reason the piece of evidence was unacceptable. At the conclusion of this process, these notes were analyzed to identify seven categories of evidence: acceptable, miscategorized, and five reasons a piece of evidence may be deemed unacceptable. As a result of this process, seven coding labels were identified, as shown in Table 62 below. The examples in this table are taken directly from the coded data. The descriptions were written after the exercise on the initial 196 data, but before coding the remainder of the data.

Table 62: The seven categories used to code the evidence teams provided in support of their claims and connections in their models. The letters in parentheses in the left ('Coding Label') column give the abbreviations for the tables below.

Coding Label	Description	Example
Acceptable (A)	The piece of evidence is acceptable. It applies to this connection or claim, it provides reason to believe the claim is true, it is given the proper category, and it does not repeat the justifications given in any prior pieces of evidence.	Logical Explanation: "Since Algal Blooms are a plant, they consume more energy when the sun is out, work better, and do more."
Miscategorized (M)	The text of the evidence itself is acceptable, it does not repeat the justifications given in any prior pieces of evidence, and it applies to this connection or claim. However, the category given for the piece of evidence is incorrect.	Controlled Experiment: "The boats are polluting the water." Direct Observation: "High amounts of fertilizer may cause too high of amounts of algae."
Redundant (R)	The piece of evidence does not add anything to the justification for this claim or connection that was not already stated by an earlier piece of evidence.	Simulation Observations: "The fish won't get enough oxygen." and "The fish are not breathing in enough oxygen."
Gibberish (G)	The text of this piece of evidence is either missing or indecipherable.	Direct Observation: "Dash Riptide a commenter said 'the PPC was very happy about their own chicken litter.'"
Irrelevant (I)	While this piece of evidence provides information that could be used to justify some claim, it does not apply in any way to the claim it is being used to justify here.	Expert Information: "In 2004, the U.S. reported to have produced 19% of the world's energy-related carbon emissions." as evidence that rising emissions cause global warming."
Insufficient (S)	While this piece of evidence suggests something that could be used in defense of this claim or connection, it is not sufficient on its own.	Logical Explanation: "Toxic algae will cause oxygen levels to go down."
Not Evidence (N)	The statement does not provide any kind of evidence that could be applied to any claim or connection.	Logical Explanation: "There is really no other way that this could be happening."

In discussion, this coding scheme may be further simplified to only differentiate acceptable and unacceptable evidence. Any evidence items coded into the first two categories (Acceptable or Miscategorized) are discussed as generally acceptable, while any items coded into the other five categories are discussed as generally unacceptable.

Intra-Rater Reliability

After establishing the above coding scheme, all 1301 pieces of evidence were processed through three rounds of coding. Each round of coding was conducted by the same coder, separated by three weeks. In each round, the order of the 640 connections was randomized. All identifying information for each individual connection was hidden; only the category, text, and connection associated with each piece of evidence was visible during coding. One piece of evidence in the list included information that revealed that the evidence must have originated from a team in the Experimental group ("like the guide said"), but this was ultimately coded as Irrelevant to its connection. For those pieces of evidence coded as 'miscategorized', a new category was also included with the code.

The tables below give the results of the three rounds of coding in pairs (T.5).

Table 63: Coding agreement between the first and second round of evidence coding. Each cell represents the number of items coded with that combination of categories in the first and second round. For example, the top left cell shows the number of items both rounds marked 'Acceptable'. The second cell in the top row shows the number of items that round 1 marked 'Acceptable' and round 2 marked 'Gibberish'.

	Round 2								
Round 1		**A**	**M**	**R**	**G**	**I**	**S**	**N**	**Total**
	A	**834**	15	10	1	6	4	6	**876**
	M	2	**138**	10	0	1	2	0	**153**
	R	1	7	**45**	0	1	0	2	**56**
	G	2	0	3	**42**	1	0	5	**53**
	I	2	4	3	0	**60**	0	6	**75**
	S	1	4	1	1	7	**14**	3	**31**
	N	2	0	3	1	9	2	**40**	**57**
	Total	**844**	**168**	**75**	**45**	**85**	**0**	**62**	**1301**
Observed Agreements: 90.2% (1173/1301) Cohen's Kappa: 0.817									

Table 64: Coding agreement between the first and third round of evidence coding. See Table 63 for notes on reading the table.

		Round 3							
Round 1		**A**	**M**	**R**	**G**	**I**	**S**	**N**	**Total**
	A	**831**	18	13	1	5	3	5	**845**
	M	6	**135**	10	0	1	1	0	**48**
	R	3	6	**46**	0	0	0	1	**90**
	G	1	0	4	**43**	2	1	2	**165**
	I	2	8	5	0	**51**	0	9	**59**
	S	2	3	1	1	8	**13**	3	**66**
	N	0	0	2	2	9	3	**41**	**28**
	Total	**845**	**170**	**81**	**47**	**76**	**21**	**61**	**1301**
Observed Agreements: 89.2% (1160/1301) Cohen's Kappa: 0.798									

Table 65: Coding agreement between the second and third round of evidence coding. See Table 63 for notes on reading the table.

		Round 3							
Round 2		**A**	**M**	**R**	**G**	**I**	**S**	**N**	**Total**
	A	**828**	4	7	1	1	1	2	**844**
	M	1	**156**	9	0	2	0	0	**168**
	R	9	5	**60**	0	0	0	1	**75**
	G	1	0	0	**43**	1	0	0	**45**
	I	2	5	3	0	**69**	0	6	**85**
	S	2	0	1	0	1	**16**	2	**22**
	N	2	0	1	3	2	4	**50**	**62**
	Total	**845**	**170**	**81**	**47**	**76**	**21**	**61**	**1301**
Observed Agreements: 93.9% (1222/1301) Cohen's Kappa: 0.890									

Agreement between each pair of round of coding was regarded as very good. In most cases of disagreement, the tie could be broken between the three rounds; typically, one code would appear for a given item in two rounds with the third round disagreeing. In the twelve instances where all three rounds disagreed, the third round's code was used as the final code. Similar assessments of intra-rater reliability were used to assess coding of the correct categories for those evidence items marked as 'miscategorized' with nearly universal agreement and no three-way disagreements.

Following the intrarater reliability run above, a second round of reliability testing was conducted to examine for the more critical distinction in this coding analysis. As referenced above, pieces of evidence coded 'Acceptable' or 'Miscategorized' are considered generally acceptable because the text of the evidence supports the claims made by the team in the model. Pieces of evidence coded with any of the other categories are considered unacceptable because they do not add any real support for the teams' models. Thus, a follow-up round of reliability testing examined specifically agreement over what evidence was considered acceptable and what was considered unacceptable.

Table 66: Coding agreement between the first, second, and third rounds of coding examining only whether or not individual pieces of evidence were ruled acceptable (A) or unacceptable (U).

		2		**3**	
		A	**U**	**A**	**U**
1	**A**	989	40	990	39
	U	23	249	25	247

1-2 Agreement: 95.2%
Cohen's Kappa: 0.857

1-3 Agreement: 95.1%
Cohen's Kappa: 0.854

3	**A**	989	23
	U	26	263

2-3 Agreement: 96.2%
Cohen's Kappa: 0.891

This analysis came to greater consensus for which pieces of evidence were considered acceptable and which were considered unacceptable; a significant portion of the disagreement during the original coding process was disagreement over why a piece of evidence was unacceptable rather than whether it was unacceptable. The immediately following analysis of the coding process itself will examine the full range of coding categories, but when re-grading the models based on the results of this coding process, only the distinction between acceptable and unacceptable and the newly-assigned categories for miscategorized evidence will be used.

Overall Coding Results

In order to better understand the quality of the evidence supplied by teams as parts of their models, the results of the coding process were compared between the Control and Experimental group. Table 67 below shows the results of this comparison.

Table 67: Results of the coding process across Control and Experimental conditions, expressed as percentages of all pieces of evidence given within those groups. For reference, the categories are: A=Acceptable, M=Miscategorized, G=Gibberish, I=Irrelevant, N=Not Evidence, R=Redundant, S=Insufficient.

	A	M	G	I	N	R	S
Control	57.32%	12.20%	5.49%	8.33%	6.91%	7.93%	1.83%
Experimental	70.70%	12.48%	2.35%	5.19%	3.34%	4.57%	1.36%

To establish the significance of these differences, a χ^2 analysis was run using the percentages from the Control group as points of comparison for the Experimental group. Table 68 below provides the observed and expected numbers for this analysis.

Table 68: Observed and expected counts of evidence coded into each category in the Experimental group, as predicted by the percentages of evidence coded into each category in the Control group.

	A	M	G	I	N	R	S
Control %	57.32%	12.20%	5.49%	8.33%	6.91%	7.93%	1.83%
Experimental Expected	464	99	44	67	56	64	15
Experimental Observed	572	101	19	42	27	37	11

X^2 analysis reveals $\chi^2 = 120.512$ ($p \approx 0.0$), marking a statistically significant difference in the results between the Control and Experimental groups (T.6). This analysis shows that **teams in the Experimental group generated better evidence than teams in the Control group**. A significantly greater percentage of their evidence was marked as Acceptable. The rates of Miscategorized and Insufficient evidence between the Control and Experimental groups were relatively similar. The increase in the number of

Acceptable pieces of evidence in the Experimental group came from a drop in Gibberish, Irrelevant, Redundant, and 'Not Evidence' pieces of evidence.

This analysis may be further decomposed into individual analysis of the results in the Learning project and the Transfer project. The two tables below show the results of this coding process within each project specifically.

Table 69: Observed and expected counts of evidence coded into each category in the Experimental group during the Learning project, as predicted by the percentages of evidence coded into each category in the Control group. For reference, the categories are: A=Acceptable, M=Miscategorized, G=Gibberish, I=Irrelevant, N=Not Evidence, R=Redundant, S=Insufficient.

	A	M	G	I	N	R	S
Control	59.00%	12.64%	5.75%	8.43%	6.13%	5.36%	2.68%
Experimental	72.83%	12.61%	1.96%	5.43%	1.96%	4.57%	0.65%
Experimental Expected	271	27	39	58	28	25	12
Experimental Observed	335	9	25	58	9	21	3

Table 70: Observed and expected counts of evidence coded into each category in the Experimental group during the Transfer project, as predicted by the percentages of evidence coded into each category in the Control group.

	A	M	G	I	N	R	S
Control	55.41%	11.69%	5.19%	8.23%	7.79%	10.82%	0.87%
Experimental	67.91%	12.32%	2.87%	4.87%	5.16%	4.58%	2.29%
Experimental Expected	193	18	29	41	27	38	3
Experimental Observed	237	10	17	43	18	16	8

X^2 analysis using the Control values to predict the expected Experimental values within each project revealed significant differences in both. In the Learning project, $\chi^2 = 52.423$ ($p \approx 0.0$) (T.6L), while in the Transfer project, $\chi^2 = 42.720$ ($p \approx 0.0$) (T.6T).

Additional analysis also examined the change in evidence quality between the Learning project and the Transfer project, within both the Control group and the Experimental group, as shown by Table 71 below.

Table 71: Observed and expected counts of evidence coded into each category in the Experimental group during the Transfer project, as predicted by the percentages of evidence coded into each category in the Learning project. For reference, the categories are: A=Acceptable, M=Miscategorized, G=Gibberish, I=Irrelevant, N=Not Evidence, R=Redundant, S=Insufficient.

	A	M	G	I	N	R	S
Control % in Learning Project	59.00%	12.64%	5.75%	8.43%	6.13%	5.36%	2.68%
Transfer Project Expected	136	29	13	20	14	13	6
Transfer Project Observed	128	12	19	27	18	25	2

Within the Control group, a statistically significant change in performance was observed ($\chi^2 = 15.62$, $p < 0.05$). The changes, however, varied; the rate of Miscategorized evidence was much lower in the Transfer project than the Learning project would have predicted, but notable rises occurred in the incidence of Gibberish, Irrelevant, and Redundant evidence. Overall, it would appear that teams in the Control condition experienced a statistically significant drop in performance between the Learning and Transfer projects with a moderate effect size. This is notable because for those teams in the Control condition, the only differences between the Learning and Transfer project was the different phenomenon to investigate and the presence of a shorter time constraint.

Table 72: Observed and expected counts of evidence coded into each category in the Experimental group during the Transfer project, as predicted by the percentages of evidence coded into each category in the Learning project.

	A	M	G	I	N	R	S
Experimental % in Learning Project	72.83%	12.61%	1.96%	5.43%	1.96%	4.57%	0.65%
Transfer Project Expected	254	7	19	44	7	16	2
Transfer Project Observed	237	10	17	43	18	16	8

As shown in Table 72, Teams in the Experimental condition experienced a larger drop in performance between the Learning and Transfer projects ($\chi^2 = 37.94$, $p \approx 0.0$). Although the incidence of Miscategorized, Gibberish, Irrelevant, and Redundant evidence stayed consistent across the two projects, the prevalence of 'Not Evidence' and Insufficient evidence jumped, leading to a notable drop in the number of Acceptable pieces of evidence relative to the total number of pieces of evidence supplied. Although this would appear to be expected given the removal of a source of feedback for teams in the Experimental condition, the tutors of MILA–T cannot actually read and give feedback on the evidence that teams write; thus, in the Learning project, teams in the Experimental condition did not receive direct feedback from MILA–T items that would have otherwise been marked as 'Insufficient' or 'Not Evidence'. If they nonetheless demonstrated superior performance in these categories when MILA–T was available, this lends credence to the notion that one of the main effects of MILA–T was enabling teachers to give better feedback in the classroom. Alternatively, it may also support the idea that the superior evidence during the Learning project was due to a Hawthorne effect (Cook 1962). These ideas are explored in more detail in the discussion of the primary claim in Chapter 6.

In addition to comparing the results from the Control and Experimental teams within each project and comparing the results within each condition across projects, similar analysis was also conducted to examine the differences between teachers and class levels. Using teams in Teacher A's classes as the predictor for teams in Teacher B's classes, a statistically significant relationship was observed ($\chi^2 = 15.62$, $p < 0.05$) suggesting that Teacher B's classes experienced a higher incidence of Redundant evidence, among other minor differences. Similarly, a statistically significant relationship was observed ($\chi^2 = 16.71$, $p < 0.05$) when using results from On-Level classes to predict the results of Gifted classes. Teams in Gifted classes demonstrated more Acceptable pieces of evidence than predicted by the On-Level classes, as well as fewer Gibberish, Miscategorized, and Redundant pieces of evidence.

Summary of Coding Results

The χ^2 analysis shows that **teams in the Experimental group used evidence to justify their claims better than teams in the Control group**, specifically by writing less irrelevant, redundant, and nonsensical justifications. A number of different mechanisms can be posited to explain this difference. First, built into the feedback of the tutors of MILA–T (especially the Guide and the Mentor) was a description of what evidence should be. The tutors informed teams that evidence can be thought of as the reason why you believe your claims are true, or what you would use to convince someone else of your claims. Similarly but importantly, these tutors also provided feedback to teams encouraging them to think of their models as claims to be defended. This reframing of the modeling exercise as an opportunity to defend and debate teams' answers may have improve their understanding of the role of evidence, thus leading to the drop in observed instances of evidence coded as 'Not Evidence' or Irrelevant.

More directly, in some instances, MILA–T was equipped with the ability to directly comment on evidence that would have fallen into one of these categories, the 'Gibberish' category. MILA–T was able to detect instances where the evidence field was left blank or was very short and comment accordingly, potentially directly limiting the number of gibberish responses. The Critic and the Mentor were also both constructed to alert teams when they found themselves relying too heavily on certain "weaker" categories of evidence, especially Logical Explanations. This feedback was present both at the level of individual claims (e.g., the only defense given for a specific claim is that it makes sense) and for the model as a whole (e.g., the model as a whole seems based in observations rather than established theories). This feedback may have directly limited the incidence of redundancy in teams' evidential justifications by encouraging them to vary the types of data they sought.

These mechanisms infer that the improved use of evidence in Experimental teams is based on feedback received from MILA–T. However, other explanations exist as well.

It may be the case that the presence of MILA–T did not directly affect teams' understanding of evidence, but rather that MILA–T was capable of providing low-level feedback on model size and complexity that typically a teacher would have to provide. The teachers in the class, then, may not have had to constantly monitor low-level details of teams' progress, freeing their time to give teams' higher-level feedback on the overall quality of their evidence. In other words, the gains seen here may have come from the teachers rather than MILA–T; under this idea, MILA–T's role was in providing teachers more time to observe high-level trends and examine teams' models more deeply. The analysis of the drop-off between the Learning and Transfer project in the Experimental group lends some support to this idea: when MILA–T was disabled, teams in the Experimental condition demonstrated a higher tendency to commit mistakes that MILA–T was not equipped to address in the first place. Therefore, it may stand to reason that the presence of MILA–T actually provided the teachers with the time and cognitive resources to focus on bigger, deeper questions, leading to better performance on metrics seemingly outside MILA–T's influence. The following section, Teacher Feedback, sheds more light on this potential mechanism behind the improved understanding of evidence seen in the Experimental groups.

Another alternate explanation may describe the results more cynically as effects solely of the superficial or hygiene factors of the tutoring system rather than effects of the actual feedback provided. For example, while teams in the Experimental group did not receive feedback from the tutoring system directly on the evidence that they wrote (given the lack of natural language processing in the tutoring system), they may have nonetheless felt as if there statements and decisions there were being witnessed by MILA–T. This perceived observation of their decisions and text may have itself caused teams to put more effort into the evidence that they wrote as an instance of the Hawthorne effect (Cook 1962). Similarly, the tutors of MILA–T were equipped with faces and facial expressions that differed based on the decisions that teams made; this, in turn, may have reinforced this feeling of being observed and further improved teams' performance without teams even considering the specific nature of the advice received from the tutoring system. These ideas are explored further in the "Discussion of Primary Claim" section of Chapter 6 on page 216.

It is worth noting, however, that the rates of Miscategorized evidence were the same between the Control and Experimental groups. Among the potential improvements based on interaction with MILA–T, this shows a potential improvement that did not quite materialize. The tutors of MILA–T were equipped with feedback to describe and define the different kinds of evidence to teams. One might expect, then, that the rate of Miscategorized evidence would fall in the Experimental group, but that result was not observed. A possible explanation for this, however, is that while MILA–T also alerted teams to the definitions of different kinds of evidence, it also alerted teams more directly to the strengths associated with different kinds of evidence. Thus, it may have also caused teams to try to categorize evidence according to the "stronger" categories even when it was not a natural fit, leading to more instances of Miscategorized evidence and mitigating the potential improvements in understanding of the categories.

Team-Based Coding

Cursory perusal of the results of the coding of evidence suggested, however, that there existed significant variation between groups with regard to the quality of evidence provided. For example, in the Learning project, teams in the Control condition registered 154 pieces of Acceptable evidence out of 261 total; however, 33 (21%) of these pieces came from three (10%) of the Control teams. Similarly, teams in the Control condition registered 22 pieces of Irrelevant evidence on the Learning project, but 9 (41%) of these pieces came from a single one of the 31 teams. Although difficult to empirically demonstrate given the low number of pieces of evidence contributed by each team, it would stand to reason that when a team has misconceptions about the meaning and role of evidence in their model, those misconceptions would influence multiple pieces of evidence instead of influencing the team's results according to the mean of the Condition as a whole.

In order to examine the potential influence of this effect, the results of the coding process were recalculated to weight each group evenly. The proportion of pieces of evidence falling into each of the seven categories were first calculated within each individual team, and those proportions were then averaged overall. This risks the opposite effect of the previous analysis: in the previous analysis, teams with several pieces of evidence within each category weighted the overall results, while in this analysis, teams with few pieces of evidence within each category weighted the overall results. In other words, a team for which both of its two pieces of evidence were labeled Acceptable would be as influential to the final averages in this analysis as a team for which all ten of its pieces of evidence were labeled Acceptable. Thus, this analysis is not superior to the previous analysis, but rather simply provides an alternate view of the results.

The results of this compilation of the evidence are shown in Table 73 and Table 74 below.

Table 73: The reinterpreted results of the coding process for the Learning project across Control and Experimental conditions, expressed as percentages of all pieces of evidence given within those groups. For reference, the categories are: A=Acceptable, M=Miscategorized, G=Gibberish, I=Irrelevant, N=Not Evidence, R=Redundant, S=Insufficient.

	A	M	G	I	N	R	S
Control	62.04%	13.54%	4.78%	6.59%	5.09%	5.31%	2.64%
Experimental	76.07%	12.04%	1.52%	4.69%	1.59%	3.38%	0.72%

Table 74: The reinterpreted results of the coding process for the Transfer project across Control and Experimental conditions, expressed as percentages of all pieces of evidence given within those groups.

	A	M	G	I	N	R	S
Control	49.31%	16.75%	6.87%	6.45%	10.16%	9.84%	0.62%
Experimental	71.36%	12.26%	2.36%	3.18%	3.76%	3.74%	3.34%

Based on this reinterpretation, the conclusions of the previous analysis remain the same: **teams in the Experimental group outperformed teams in the Control group even when interpreting the quality of evidence at the model level rather than at the level of individual pieces of evidence**. Notably, however, in the Learning project, teams in both the Control group and the Experimental group experienced an increase in the rate of Acceptable evidence and a decrease across several categories of unacceptable evidence (Gibberish, Irrelevant, and 'Not Evidence' specifically). This suggests that in the original results, a small number of teams were having a disproportionately high effect on the results of the coding, such as the previous example of a single team accounting for 9 of the 22 pieces of irrelevant evidence in the Learning project.

In the Transfer project, a different effect was observed. Results for teams in the Experimental condition improved under this reinterpretation, with a modest increase in Acceptable evidence and decreases in evidence coded Irrelevant or 'Not Evidence'. This suggests again that the instances of unacceptable evidence were contained within a small number of teams. However, for teams in the Control condition, the opposite effect took place: the rate of Acceptable evidence actually dropped while the rate of Miscategorized evidence rose, along with the rate of unacceptable evidence coded as Gibberish or 'Not Evidence'. This suggests that in the Control condition, a disproportionately small number of teams were responsible for the instances of Acceptable evidence. For example, on the Transfer project in the Control condition, of 128 pieces of Acceptable evidence, 25% (32 out of 128) came from only three (of 31, 9%) of the teams.

As mentioned previously, the results of this modified interpretation of the results of the evidence coding process suggest the same conclusions as the prior analysis; in both the Learning and the Transfer project, teams in the Experimental condition demonstrated an increased incidence of evidence coded Acceptable and a decreased incidence of evidence coded into any of the unacceptable categories of evidence, whether summarized at the evidence level or the model level.

Coded Model Analysis

The above process of evidence coding significantly alters the scoring of teams' models according to the previous metrics. Although nothing in the coding process would alter the complexity of teams' models, significant portions of this process may alter the strength of evidence that teams provide. In order to identify this effect, the metrics for each teams' models were recalculated according to the results of this coding process. Any evidence coded as redundant, gibberish, irrelevant, insufficient, or not evidence was thrown out. Any evidence marked as miscategorized was recalculated according to its new category instead of its previous category.

Learning Project

As before, analysis of the results of the Learning project began with a multivariate analysis of variance to examine the effect of the three independent variables (Condition, Teacher, and Class Level) on the five dependent variables (Model Complexity, Total Model Strength, Average Model Strength, Average Evidence Strength, and Total Evidence). The process of evidence coding and model regrading did not affect the Model Complexity statistic, and thus while it is included in this analysis to more directly compare the results to the previous analysis, no changes will be observed in it.

Initial multivariate analysis of variance revealed a statistically significant effect of Condition (F = 3.30, $p < 0.01$), Teacher (F = 2.54, $p < 0.05$), and Class Level (F = 50, $p \approx 0.0$) (T.7L). The results of the initial analysis of variance are shown below in Table 75.

Table 75: Multivariate ANOVA results on the differences in teams' models across five metrics based on different independent variables on the Learning project after the evidence coding process. The results indicate statistically significant effects of all three independent variables: Condition (F = 2.91, $p < 0.05$), Teacher (F = 3.12, $p < 0.05$), and Class Level (F = 8.21, $p \approx 0.0$)

```
                     Df  Pillai approx F num Df den Df    Pr(>F)
Condition             1 0.16452   2.9143      5     74   0.01862
Teacher               1 0.17434   3.1251      5     74   0.01295
ClassLevel            1 0.35684   8.2114      5     74 3.291e-06
Condition:Teacher     1 0.04268   0.6598      5     74   0.65505
Condition:ClassLevel  1 0.05791   0.9098      5     74   0.47951
Residuals            78
```

Table 76: Univariate ANOVA results each of the five metrics describing the differences in teams' models across the different independent variables after the evidence coding and model regrading.

```
 Response P1.TotalStrength :
                     Df Sum Sq Mean Sq F value   Pr(>F)
Condition             1  746.7  746.68 10.8714 0.001472
Teacher               1  277.6  277.57  4.0413 0.047856
ClassLevel            1  778.1  778.13 11.3292 0.001186
Condition:Teacher     1   28.4   28.42  0.4138 0.521909
Condition:ClassLevel  1   24.2   24.23  0.3528 0.554253
Residuals            78 5357.3   68.68

 Response P1.AverageStrength :
                     Df  Sum Sq Mean Sq F value   Pr(>F)
Condition             1  23.268  23.268  7.4604 0.007796
Teacher               1  35.517  35.517 11.3878 0.001154
ClassLevel            1   2.725   2.725  0.8738 0.352798
Condition:Teacher     1   1.013   1.013  0.3248 0.570361
Condition:ClassLevel  1   0.038   0.038  0.0122 0.912233
Residuals            78 243.267   3.119

 Response P1.TotalEvidence :
                     Df  Sum Sq Mean Sq F value   Pr(>F)
Condition             1  112.72 112.718  5.9196 0.017263
Teacher               1   98.72  98.722  5.1845 0.025530
ClassLevel            1  188.82 188.818  9.9161 0.002322
Condition:Teacher     1   13.08  13.084  0.6871 0.409681
Condition:ClassLevel  1    2.65   2.653  0.1393 0.709987
Residuals            78 1485.24  19.042

 Response P1.AverageEvidenceStrength :
                     Df  Sum Sq Mean Sq F value   Pr(>F)
Condition             1  1.4589 1.45887  3.8004 0.054834
Teacher               1  0.4403 0.44034  1.1471 0.287460
ClassLevel            1  2.9416 2.94159  7.6630 0.007038
Condition:Teacher     1  0.0617 0.06171  0.1608 0.689559
Condition:ClassLevel  1  0.0131 0.01315  0.0342 0.853661
Residuals            78 29.9419 0.38387
```

Based on the results of this analysis, a follow-up univariate analysis of variance was conducted on each of the four altered metrics. The raw results of this univariate analysis are shown above in Table 76. Compared with analysis of these data prior to the evidence coding and model regrading, several differences remain, but other conclusions change. First, as in the original analysis, a statistically significant effect of Condition is observed on Total Model Strength and Average Model Strength. The prior difference in Average Evidence Strength has dropped slightly and is no longer statistically significant (previously F = 4.70, *p* = 0.033; now, F = 3.80, *p* = 0.054), but in its stead there is now a statistically significant difference in Total Evidence (previously F = 1.96, p = 0.166; now, F = 5.92, p = 0.017) (T.7L.A-E). This matches the results of the χ^2 analysis of the evidence coding itself: teams in the Control group experienced a higher incidence of evidence coded under unacceptable categories, and in this regarding, these pieces of evidence were thrown out. The inclusion of a greater percentage of Experimental group's evidence leads to the difference in Total Evidence between the two groups. The change in Average Evidence Strength is interesting, but ultimately the actual change between the earlier analysis and this analysis is likely not significant enough for close inspection.

As before, the charts below report the numerical results of the analysis of the Learning models after evidence coding and model regrading. Given that evidence does not affect the Model Complexity calculation, the chart reporting results within the Model Complexity metric is not duplicated here. It can be seen in Table 50 on page 158.

Table 77: Difference in Total Model Strength after the evidence coding process between teams in different Conditions, Teachers, Class Levels, and combinations thereof. The top number in each cell is the mean for that group; the number in parentheses is the standard deviation within that group; the bottom number is the number of samples within that subgroup. See Table 47 for further information about the groups of students summarized in each cell.

Difference in Total Model Strength in the Learning Project after Evidence Coding

<table>
<tr><td colspan="2">Teacher A</td><td colspan="2">Teacher B</td><td colspan="2">Teacher A</td><td colspan="2">Teacher B</td></tr>
<tr><td colspan="2">7.19
(5.74)
<i>n</i> = 16</td><td colspan="2">9.67
(8.03)
<i>n</i> = 18</td><td colspan="2">11.88
(10.82)
<i>n</i> = 25</td><td colspan="2">16.68
(9.33)
<i>n</i> = 25</td></tr>
<tr><td colspan="4">Control</td><td colspan="4">Experimental</td></tr>
<tr><td colspan="4">8.50
(7.05)
<i>n</i> = 34</td><td colspan="4">14.28
(10.29)
<i>n</i> = 50</td></tr>
<tr><td colspan="2">On-Level</td><td colspan="2">Gifted</td><td colspan="2">On-Level</td><td colspan="2">Gifted</td></tr>
<tr><td colspan="2">7.82
(7.19)
<i>n</i> = 28</td><td colspan="2">11.67
(5.92)
<i>n</i> = 6</td><td colspan="2">13.38
(9.87)
<i>n</i> = 42</td><td colspan="2">19.00
(11.82)
<i>n</i> = 8</td></tr>
<tr><td>A</td><td>B</td><td>A</td><td>B</td><td>A</td><td>B</td><td>A</td><td>B</td></tr>
<tr><td>4.50
(3.75)
<i>n</i> = 10</td><td>9.67
(8.03)
<i>n</i> = 18</td><td>11.67
(5.92)
<i>n</i> = 6</td><td>--
--
<i>n</i> = 0</td><td>8.53
(8.78)
<i>n</i> = 17</td><td>16.68
(9.33)
<i>n</i> = 25</td><td>19.00
(11.82)
<i>n</i> = 8</td><td>--
--
<i>n</i> = 0</td></tr>
</table>

Table 78: Difference in Average Model Strength after the evidence coding process between teams in different Conditions, Teachers, Class Levels, and combinations thereof. The top number in each cell is the mean for that group; the number in parentheses is the standard deviation within that group; the bottom number is the number of samples within that subgroup. See Table 47 for further information about the groups of students summarized in each cell.

Difference in Average Model Strength in the Learning Project after Evidence Coding

Teacher A		Teacher B		Teacher A		Teacher B	
1.40 (1.07) *n* = 16		2.41 (1.89) *n* = 18		2.26 (1.61) *n* = 25		3.75 (2.10) *n* = 25	
Control				Experimental			
1.94 (1.62) *n* = 34				3.01 (2.00) *n* = 50			
On-Level		Gifted		On-Level		Gifted	
1.97 (1.70) *n* = 28		1.78 (1.30) *n* = 6		3.09 (2.13) *n* = 42		2.58 (1.01) *n* = 8	
A	B	A	B	A	B	A	B
1.17 (0.91) *n* = 10	2.41 (1.89) *n* = 18	1.78 (1.30) *n* = 6	-- -- *n* = 0	2.11 (1.83) *n* = 17	3.75 (2.10) *n* = 25	2.58 (1.01) *n* = 8	-- -- *n* = 0

Table 79: Difference in Average Evidence Strength after the evidence coding process between teams in different Conditions, Teachers, Class Levels, and combinations thereof. The top number in each cell is the mean for that group; the number in parentheses is the standard deviation within that group; the bottom number is the number of samples within that subgroup. See Table 47 for further information about the groups of students summarized in each cell.

Difference in Average Evidence Strength in the Learning Project after Coding

Teacher A		Teacher B		Teacher A		Teacher B	
1.25 (0.60) *n* = 16		1.45 (0.59) *n* = 18		1.57 (0.81) *n* = 25		1.68 (0.5) *n* = 25	
Control				Experimental			
1.36 (0.59) *n* = 34				1.62 (0.67) *n* = 50			
On-Level		Gifted		On-Level		Gifted	
1.29 (0.61) *n* = 28		1.64 (0.43) *n* = 6		1.57 (0.70) *n* = 42		1.93 (0.32) *n* = 8	
A	B	A	B	A	B	A	B
1.02 (0.58) *n* = 10	1.45 (0.59) *n* = 18	1.64 (0.43) *n* = 6	-- -- *n* = 0	1.40 (0.91) *n* = 17	1.68 (0.50) *n* = 25	1.93 (0.32) *n* = 8	-- -- *n* = 0

Table 80: Difference in Total Evidence after the evidence coding process between teams in different Conditions, Teachers, Class Levels, and combinations thereof. The top number in each cell is the mean for that group; the number in parentheses is the standard deviation within that group; the bottom number is the number of samples within that subgroup. See Table 47 for further information about the groups of students summarized in each cell.

Difference in Total Evidence in the Learning Project after Evidence Coding

Teacher A		Teacher B		Teacher A		Teacher B	
4.94 (3.66) *n* = 16		6.00 (3.40) *n* = 18		6.4 (5.37) *n* = 25		9.32 (4.93) *n* = 25	
Control				Experimental			
5.50 (3.51) *n* = 34				7.86 (5.31) *n* = 50			
On-Level		Gifted		On-Level		Gifted	
5.11 (3.42) *n* = 28		7.33 (3.61) *n* = 6		7.50 (5.26) *n* = 42		9.75 (5.50) *n* = 8	
A	B	A	B	A	B	A	B
3.50 (2.99) *n* = 10	6.00 (3.40) *n* = 18	7.33 (3.61) *n* = 6	-- -- *n* = 0	4.82 (4.67) *n* = 17	9.32 (4.93) *n* = 25	9.75 (5.50) *n* = 8	-- -- *n* = 0

Transfer Project

The multivariate analysis of variance was then again conducted on the models in the Transfer project after evidence coding and model regrading. As before, the goal here is to discern the effect of the three independent variables (Condition, Teacher, and Class Level) on the five dependent variables (Model Complexity, Total Model Strength, Average Model Strength, Average Evidence Strength, and Total Evidence). Similarly, as before, the Model Complexity status is included in the analysis for consistency, but remains unchanged by the coding process.

Table 81: Multivariate ANOVA results on the differences in teams' models across five metrics based on different independent variables. The results indicate statistically significant effects of two independent variables: Condition (F = 4.23, *p* < 0.01) and Teacher (F = 3.90, *p* < 0.01).

```
                     Df   Pillai  approx F  num Df  den Df    Pr(>F)
Condition             1 0.229323    4.2254       5      71 0.002029
Teacher               1 0.215394    3.8982       5      71 0.003527
ClassLevel            1 0.092597    1.4491       5      71 0.217509
Condition:Teacher     1 0.077698    1.1963       5      71 0.319863
Condition:ClassLevel  1 0.027497    0.4015       5      71 0.846227
Residuals            75
```

Initial multivariate analysis of variance revealed a statistically significant effect of Condition (F = 4.23, *p* < 0.01) and Teacher (F = 3.90, *p* < 0.01), but not from Class Level (T.7T). The results of the multivariate analysis of variance are shown above in Table 81.

Table 82: Univariate ANOVA results each of the five metrics describing the differences in teams' models across the different independent variables after the evidence coding and model regrading.

```
Response P2.TotalStrength :
                     Df Sum Sq Mean Sq F value    Pr(>F)
Condition             1  509.1  509.14 12.0024 0.0008819
Teacher               1  186.9  186.94  4.4069 0.0391584
ClassLevel            1  119.1  119.12  2.8082 0.0979491
Condition:Teacher     1    0.3    0.29  0.0067 0.9348601
Condition:ClassLevel  1   81.0   81.03  1.9101 0.1710530
Residuals            75 3181.5   42.42

Response P2.AverageStrength :
                     Df  Sum Sq Mean Sq F value  Pr(>F)
Condition             1  15.208 15.2081  5.9675 0.01693
Teacher               1   0.534  0.5339  0.2095 0.64849
ClassLevel            1   0.321  0.3207  0.1259 0.72376
Condition:Teacher     1   0.173  0.1728  0.0678 0.79526
Condition:ClassLevel  1   2.550  2.5503  1.0007 0.32036
Residuals            75 191.138  2.5485

Response P2.TotalEvidence :
                     Df Sum Sq Mean Sq F value  Pr(>F)
Condition             1  38.59  38.592  4.1542 0.04506
Teacher               1  42.04  42.043  4.5257 0.03668
ClassLevel            1  24.43  24.429  2.6297 0.10908
Condition:Teacher     1   1.16   1.164  0.1253 0.72433
Condition:ClassLevel  1  15.92  15.919  1.7136 0.19452
Residuals            75 696.74   9.290

Response P2.AverageEvidenceStrength :
                     Df Sum Sq Mean Sq F value    Pr(>F)
Condition             1  6.384  6.3838 12.2128 0.0008003
Teacher               1  0.543  0.5431  1.0390 0.3113386
ClassLevel            1  1.136  1.1362  2.1736 0.1445824
Condition:Teacher     1  0.020  0.0205  0.0392 0.8436485
Condition:ClassLevel  1  0.321  0.3210  0.6141 0.4357062
Residuals            75 39.204  0.5227
```

Based on these results, a follow-up univariate analysis of variance was conducted on each of the five metrics. The raw results of this univariate analysis are shown above in Table 82, and the raw numbers observed across the four metrics in the analysis after evidence coding and regrading are shown in Table 83, Table 84, Table 85, and Table 86. The evidence coding and model regrading process maintained many of the results from the earlier analysis of teams' models: there remain statistically significant differences in Total Model Strength and Average Evidence Strength between the Control and Experimental groups. However, this analysis introduces two additional effects of the Condition: there now also exist statistically significant effects of the Condition on both Average Model Strength (previously F = 0.87, *p* = 0.35; now F = 5.97, *p* = 0.017) and Total Evidence (previously F = 0.66, *p* = 0.418; now F = 4.15, *p* = 0.045) (T.7T.A-E). Both these new results match with the results of the earlier χ^2 analysis of the results of the evidence coding process, which saw a much higher proportion of evidence from the Experimental group coded as Acceptable than in the Control group. This analysis also averages models at the model level rather than at the level of individual pieces of evidence, and thus the changes are further explained by the Team-Based Coding process, which saw a small number of low-performing teams lower the Experimental group's results, while a small number of high-performing teams raised the Control group's results.

Table 83: Difference in Total Model Strength between teams in different Conditions, Teachers, Class Levels, and combinations thereof. The top number in each cell is the mean for that group; the number in parentheses is the standard deviation within that group; the bottom number is the number of samples within that subgroup. See Table 47 for further information about the groups of students summarized in each cell.

Difference in Total Model Strength in the Transfer Project after Evidence Coding

Teacher A		Teacher B		Teacher A		Teacher B	
8.69 (6.68) *n* = 16		5.39 (4.39) *n* = 18		13.36 (7.39) *n* = 25		10.50 (7.14) *n* = 22	
Control				Experimental			
6.94 (5.75) *n* = 34				12.02 (7.34) *n* = 47			
On-Level		Gifted		On-Level		Gifted	
5.61 (4.45) *n* = 28		13.17 (7.39) *n* = 6		11.59 (7.65) *n* = 39		14.12 (5.49) *n* = 8	
A	B	A	B	A	B	A	B
6.00 (4.76) *n* = 10	5.39 (4.39) *n* = 18	13.17 (7.39) *n* = 6	-- -- *n* = 0	13.00 (8.27) *n* = 17	10.50 (7.14) *n* = 22	14.12 (5.49) *n* = 8	-- -- *n* = 0

Table 84: Difference in Average Model Strength between teams in different Conditions, Teachers, Class Levels, and combinations thereof. The top number in each cell is the mean for that group; the number in parentheses is the standard deviation within that group; the bottom number is the number of samples within that subgroup. See Table 47 for further information about the groups of students summarized in each cell.

Difference in Average Model Strength in the Transfer Project after Evidence Coding

Teacher A		Teacher B		Teacher A		Teacher B	
1.72 (1.24) *n* = 16		1.76 (1.26) *n* = 18		2.50 (1.31) *n* = 25		2.75 (2.22) *n* = 22	
Control				Experimental			
1.74 (1.23) *n* = 34				2.62 (1.77) *n* = 47			
On-Level		Gifted		On-Level		Gifted	
1.64 (1.19) *n* = 28		2.23 (1.39) *n* = 6		2.68 (1.90) *n* = 39		2.33 (1.03) *n* = 8	
A	B	A	B	A	B	A	B
1.41 (1.10) *n* = 10	1.76 (1.26) *n* = 18	2.23 (1.39) *n* = 6	-- -- *n* = 0	2.59 (1.44) *n* = 17	2.75 (2.22) *n* = 22	2.33 (1.03) *n* = 8	-- -- *n* = 0

Table 85: Difference in Average Evidence Strength between teams in different Conditions, Teachers, Class Levels, and combinations thereof. The top number in each cell is the mean for that group; the number in parentheses is the standard deviation within that group; the bottom number is the number of samples within that subgroup. See Table 47 for further information about the groups of students summarized in each cell.

Difference in Average Evidence Strength in the Transfer Project after Coding

Teacher A		Teacher B		Teacher A		Teacher B	
1.41 (0.57) *n* = 16		1.20 (0.68) *n* = 18		1.93 (0.69) *n* = 25		1.80 (0.89) *n* = 22	
Control				Experimental			
1.30 (0.63) *n* = 34				1.87 (0.78) *n* = 47			
On-Level		Gifted		On-Level		Gifted	
1.20 (0.64) *n* = 28		1.77 (0.34) *n* = 6		1.83 (0.82) *n* = 39		2.06 (0.56) *n* = 8	
A	B	A	B	A	B	A	B
1.19 (0.58) *n* = 10	1.20 (0.68) *n* = 18	1.77 (0.34) *n* = 6	-- -- *n* = 0	1.86 (0.75) *n* = 17	1.80 (0.89) *n* = 22	2.06 (0.56) *n* = 8	-- -- *n* = 0

Table 86: Difference in Total Evidence between teams in different Conditions, Teachers, Class Levels, and combinations thereof. The top number in each cell is the mean for that group; the number in parentheses is the standard deviation within that group; the bottom number is the number of samples within that subgroup. See Table 47 for further information about the groups of students summarized in each cell.

Difference in Total Evidence in the Transfer Project after Evidence Coding

<table>
<tr><td colspan="2">Teacher A</td><td colspan="2">Teacher B</td><td colspan="2">Teacher A</td><td colspan="2">Teacher B</td></tr>
<tr><td colspan="2">5.50
(3.58)
<i>n</i> = 16</td><td colspan="2">3.72
(2.78)
<i>n</i> = 18</td><td colspan="2">6.52
(2.89)
<i>n</i> = 25</td><td colspan="2">5.32
(3.18)
<i>n</i> = 22</td></tr>
<tr><td colspan="4">Control</td><td colspan="4">Experimental</td></tr>
<tr><td colspan="4">4.56
(3.26)
<i>n</i> = 34</td><td colspan="4">5.96
(3.06)
<i>n</i> = 47</td></tr>
<tr><td colspan="2">On-Level</td><td colspan="2">Gifted</td><td colspan="2">On-Level</td><td colspan="2">Gifted</td></tr>
<tr><td colspan="2">3.93
(2.84)
<i>n</i> = 28</td><td colspan="2">7.50
(3.73)
<i>n</i> = 6</td><td colspan="2">5.77
(3.12)
<i>n</i> = 39</td><td colspan="2">6.88
(2.70)
<i>n</i> = 8</td></tr>
<tr><td>A</td><td>B</td><td>A</td><td>B</td><td>A</td><td>B</td><td>A</td><td>B</td></tr>
<tr><td>4.30
(3.06)
<i>n</i> = 10</td><td>3.72
(2.78)
<i>n</i> = 18</td><td>7.50
(3.73)
<i>n</i> = 6</td><td>--
--
<i>n</i> = 0</td><td>6.35
(3.04)
<i>n</i> = 17</td><td>5.32
(3.18)
<i>n</i> = 22</td><td>6.88
(2.70)
<i>n</i> = 8</td><td>--
--
<i>n</i> = 0</td></tr>
</table>

Summary of Output Hypothesis Results

Based on the analysis provided above, there exists sufficient evidence to accept both facets of the fourth hypothesis: **teams in the Experimental group outperform teams in the Control group in both the Learning and the Transfer projects**.

Learning Project

First, in the Learning project, teams in the Experimental group are equipped with a set of tutors that help the team with the modeling and inquiry process. The additional feedback provided by these tutors has a positive effect. In the initial analysis, teams in the Experimental group constructed stronger evidential justifications for their models than teams in the Control group. These evidential justifications were analyzed across a collection of metrics weighting certain kinds of evidence, such as citing established scientific principles and transferring explanations from similar systems, as more desirable and other kinds, like logical explanations and direct observations alone, as less desirable.

This weighting scheme is described in detail in Chapter 3. When coding their evidential justifications themselves, teams in the Experimental group constructed superior explanations based on scores calculated according to this weighting scheme.

However, implicit in this analysis was the assumption that teams were coding their evidence accurately, and moreover, writing useful evidence in the first place. It is possible that teams could have instead internalized the feedback of the tutoring system as asking the teams simply to label more of their evidence with the stronger categories rather than choose better justifications in the first place. To examine this, the evidence was processed through three rounds of coding to assess whether teams were writing sufficient justifications and coding them correctly. Here again, teams in the Experimental condition outperformed teams in the Control condition. A greater number of the pieces of evidence written by the Experimental group were coded as Acceptable (72.83%) than in the Control group (59.00%). Interestingly, there was no difference in the incidence of Miscategorized evidence between the two groups. Assessed based on the results of the evidence coding alone, however, teams in the Experimental group demonstrated an improved grasp of the nature and purpose of justifying their explanations. In many cases, these improvements came in areas that are difficult to attribute to the tutoring system, given its inability to read and interpret teams' actual evidence; this leads to the hypothesis that the main effect of the tutoring system may have been to free up time and cognitive resources in the teachers to provide better feedback as the tutoring system handles more basic, low-level questions.

After this coding process, teams models' were regraded according to the same criteria as before. Pieces of evidence that were coded as unacceptable were thrown out, and pieces of evidence that were coded as Miscategorized were calculated according to their correct category. The results of this analysis confirmed and enhanced the results of the initial analysis: teams in the Experimental condition constructed models with better evidential justifications than models constructed by teams in the Control condition. These models were superior both before the evidence coding process and after, while the evidence itself was coded as superior for teams in the Experimental condition. Based on these three facets, **Hypothesis #4A is accepted**. Teams with a metacognitive tutoring system outperformed teams without a metacognitive tutoring system on a modeling task.

Transfer Project

An argument could be made, however, that the results of the above analysis are unsurprising. In both the Control and Experimental conditions, teams had access to teachers for help, and in the Experimental condition, teams also had access to a tutoring system. Thus, teams in Experimental condition had access to more feedback, and these results could thus be summarized dismissively as teams unsurprisingly performing better when they receive more feedback. This then introduces the question: did those teams that performed better actually learn more, or were they simply performing better because they were receiving greater feedback on their work? During the Transfer project, the same teams started new projects examining a new phenomenon. During this phase, the tutoring system was disabled for teams in both the Control and the Experimental conditions. In this way, the Transfer project aimed to identify whether the improvements seen in the Learning project would transfer to a new project where the additional feedback was not available.

An initial analysis was conducted on the Transfer project the same way it was conducted on the Learning project; the results of teams' modeling processes were compared across five metrics. Here again, teams in the Experimental group outperformed teams in the Control group across on multiple metrics. Specifically, teams in the Experimental group constructed larger models (as measured by the Model Complexity variable) and again provided better evidential justifications, this time measured as providing sufficient evidence to support the larger model and using stronger individual pieces of evidence to do so.

The prior questions about the strength of teams' evidence and categorizations remained, however, and so the evidence teams wrote for the Transfer project was again coded for acceptability. Compared to the Learning project, performance for both groups dropped in the Transfer project by approximately the same amount. It would make sense that the performance of teams in the Experimental group would drop given the removal of a source of feedback they had previously received, but the presence of a similar drop in the Control group suggests that this drop may instead be attributed to the increased time constraint during the Transfer project (one class period instead of four). With the drop affecting both groups similarly, teams in the Experimental group still outperformed teams in the Control group, with a greater number of pieces of evidence from teams in the Experimental group coded as Acceptable (67.91%) than in the Control group (55.41%).

Based on the results of that evidence coding and model regrading process, the original analysis was run again, this time using the model scores derived from the results of the evidence coding process. The conclusions of this analysis confirmed and even enhanced the conclusions of the original analysis. Teams in the Experimental condition outperformed teams in the Control condition across all five metrics after the process of evidence coding and model regrading was factored into the analysis. Based on the improved performance of the Experimental group across all three of these analyses, **Hypothesis #4B** is accepted. In a new task with no tutoring system, teams who had previously participated in a modeling task with a metacognitive tutoring system outperformed teams who had participated in a modeling task without a metacognitive tutoring system.

Teacher Feedback on Tutoring System

This intervention involved the participation of five educators in the classroom: two teachers (Teacher A and Teacher B), two instructional assistants, and one supervising teacher. During the intervention, the two instructional assistants provided ongoing support to the special education students, while the supervising teacher aided with the preparation of materials prior to the intervention. Three of these individuals – the two teachers and the instructional assistant who helped Teacher A – agreed to be interviewed at the conclusion of the intervention. The full interview script is provided in Appendix D.

Interviews focused on three major topics that are also used to organize the following report on educators' experiences. First, educators were asked to reflect on MILA on its own and its strengths, weaknesses, and areas for improvement. Second, they were asked to reflect specifically on MILA–T, how it impacted those students who received it, and its strengths and weaknesses. Third, educators were asked to reflect on the overall unit including the accompanying lessons, study materials, and experiments.

Feedback on Engagement with MILA

Both teachers remarked on the ease of use of MILA and noted that teams were able to quickly become engaged with the program. In the case of Teacher A, this ran directly counter to her expectations for students in her class; she said:

> *"I was a little skeptical I'll be honest at first. I thought it was going to be hard, I thought they were going to struggle with it and I'd have to run around a lot. I was actually pleasantly surprised at how well they handled it after a brief demonstration, and I do mean brief because I didn't give them a whole lot." – Teacher A*

Teacher B echoed Teacher A's thoughts, commenting that given the goal provided by the curriculum, teams were able to easily get engaged with the software:

> *"The program easily lends itself to student engagement if you just set them up with what they need to accomplish." –Teacher B*

Teacher A went on to discuss the benefit of modeling in MILA, noting that the act of externalizing their thoughts and beliefs about the system in a structured manner forced teams to create more thoughtful explanations than they would in an unconstrained environment:

> *"I do think that the modeling itself, making the models, forced them to have to think about what they were saying. It's easy to just scribble something on a piece of paper, but when you had to actually, 'Something goes in this box, something else goes in this box', and then you have to show the connection, you have to show evidence, I think that was a lot better than just a worksheet." – Teacher A*

When asked for feedback on how MILA could be improved for future iterations, all three educators mentioned two factors: explanation of terms and semantics and instruction on low-level skills. The meaning of certain terms, like 'connections' and 'properties', were not immediately clear to all students. Explanations of these terms were available from the Guide, and log files show that several teams accessed those explanations, but there are no data to identify whether or not access to such information directly changed those students' understanding of those terms. Second, all three desired an initial phase of software instruction that focused explicitly on the interactions with MILA, such as how to add and delete components, before moving into learning how to model in MILA. A tutorial was present for those teams in the Experimental group through the Mentor, but its execution was grounded in the first phenomenon, combining software instruction with science instruction.

All three educators provided different suggestions for potential improvements to the user interface. The instructional assistant suggested pop-up or tooltip definitions for potentially foreign words like 'phenomenon' and 'abiotic'. Teacher A suggested stronger constraints on how to phrase hypotheses, such as forcing an 'if-this-then-that' phrasing. Teacher A also suggested providing greater structure to providing evidence; instead of choosing a category and writing evidence in plaintext, a structured format for the difference kinds of evidence could be present, mirroring the structured format of the models themselves. Teacher B commented on the desire for pictures, more colors, or other visual appeal to make the software more engaging to 7th grade students. She also suggested drawing a closer connection between teams' actions and the software's responses, especially with regard to improving the quality of evidence on connections. Teacher A ultimately noted that the greatest difficulty was that students' prior "learned helplessness prevails no matter what", and that her greatest difficult was in convincing students to simply push forward and try instead of awaiting guidance.

Feedback on Engagement with MILA–T

After reflecting on MILA on its own, the educators were asked to reflect on the specific role of MILA–T. First, educators commented on the difference between classes that had access to MILA–T and those classes that did not. The main observation from Teacher B was that teams without the tutors relied more heavily on her feedback:

> *"Classes without the tutors, I had to step in, they thought they were done, then I had to look real closely at what they were doing. It took more of my time." –Teacher B*

Teams in classes without the tutors had to wait on feedback from the teacher, which in turn meant that feedback arose more slowly as the teacher was only able to provide feedback to one team at a time. Additionally, the teacher had to do more work to identify the current weaknesses in teams' models, such as viewing the evidence for each claim individually, further slowing down the feedback process. This, Teacher B suggested, led to teams with tutors pushing forward with their projects more quickly:

> *“I wish all my classes had the tutors because it definitely made it easier for them to continue with their project.” –Teacher B*

The instructional assistant agreed, suggesting that students had fewer questions for the educators in those classes with access to MILA–T than in classes without access:

> *“I think they asked less questions of those of us who were walking around. ... It did kind of take it off of us a little bit, we could say, 'Hey you need to check with your tutors', so it allowed us to go straight to the program to answer the questions instead of asking us. ... They needed some clarification, but when they asked a question we could just refer them back to the tutors.” – Instructional Assistant*

This feedback focused on easing the responsibilities of the teacher and allowing educators to focus on higher-level feedback and classroom-level trends. The primary takeaway here is that a significant benefit of the presence of MILA–T may be missed by the data on students' interactions, explanations, and attitudinal changes. Rather, a significant benefit might instead fall on the teachers: in the presence of the tutoring system, the teachers may have been able to divert their attention away from many of the questions students most commonly asked (such as "Am I done yet?"). Teacher B also commented specifically on this recurring question:

> *"The tutors were very nice to have, there was a nice benefit. When they think they're done, they can just click on the tutors, or they'll pop up automatically. The time management, it definitely helped." –Teacher B*

With the tutors taking some of the responsibility off the teachers, the teachers may have been able to more fully devote themselves to other responsibilities. This idea echoes the rationale behind the division of MILA–T into five functional roles: each functional role was designed to mimic a specific role that the teacher typically plays in the classroom. With the tutors playing these roles, the teachers here may have been able to focus on providing feedback and making observations that went beyond the capabilities of the tutors. For example, the tutoring system was not equipped with content knowledge about the target system or the ability to read students' evidential justifications, and thus, the teachers were responsible for providing feedback at this level. Similarly, the teachers could track progress among multiple teams simultaneously to monitor for common misconceptions or errors. In the absence of the tutors, the teachers still provided the feedback that the tutors were providing otherwise, and thus teams still received much of the same feedback from the teacher rather than the tutors (albeit less frequently, given the teacher's limited time). Control teams thus may have compared more favorably to Experimental teams than they would have otherwise. In this instance, however, feedback from the tutors would still lower the obligation of the teachers to provide all feedback, thus allowing the teachers to focus on deeper feedback. Additional data collection on the activity of the teacher in the classroom and the nature of feedback provided to teams by the teacher would be necessary to identify how the presence of a tutoring system in a classroom exercise changes the behavior of and demand on the teacher.

Beyond feedback strictly on the difference between Control and Experimental classes, the educators also provided feedback on the specific ways in which MILA–T helped teams. The instructional assistant noted that she was not fully aware of the degree to which teams were relying on the tutors until it was disabled for the Transfer project:

> *"After the final project, I noticed that our fourth period had the tutors, and they were asking where they were, so they were actually relying on them to direct them in their final project." – Instructional Assistant*

The instructional assistant went on to comment more narrowly on the specific design of the tutors. She remarked that she appreciated that the tutors facilitated iterative and ongoing improvement by reiterating feedback that teams had not yet internalized or incorporated into their models:

> *“I liked that the tutors didn't go away until they fixed the problem. Some of the kids got annoyed by that, but I think that was a good thing, it's something we do on a daily basis.” – Instructional Assistant*

Teacher A's primary commentary on the tutors focused on the degree to which teams actually used them. She commented first that the Gifted students did not use the tutoring system at all, suggesting that gifted students' engagement with the tutoring system, while notable in the logs of student interaction, may not have been visible at the classroom level. However, her belief was that this was because these students did not need the level of feedback provided by the tutors:

> *" I know there's a few kids in those classes, they don't need the tutors, they got it, they understand what they're doing and run with it." –Teacher A*

This also reflects a misconception on the part of the teacher that tutoring is only for underperforming students, not to also keep high-performing students engaged. Among her other students, Teacher A similarly noted that many teams did not seem to use the tutoring system; however, of the teams that did, she believed there was a tangible benefit:

> *"I don't know what [students without tutors'] models looked like as a whole. I think their models might not have been as great as the [students] that had the tutors. ... Those that actually used it benefited from it, they really did." –Teacher A*

Interestingly, while Teacher B and the instructional assistant commented that the tutors helped them by taking care of many of the questions that students were asking, Teacher A commented on difficulty with the same trend. While the tutors were equipped to answer many of the students' questions, she found herself wanting to do the answering instead:

> *"I wanted to be the Guide and the Critic all the time, so I found it hard to not be the Guide and the Critic a lot." –Teacher A*

In this way, one of the benefits that Teacher B and the instructional assistant observed may have not been realized by Teacher A, who instead found it difficult to allow the tutors to provide feedback instead. This is especially interesting given that the instructional assistant, who previously had commented that the tutors meant teams asked the teachers fewer questions, worked in the same classes as Teacher A. This feedback may also explain Teacher A's comment that many teams did not seem to use the tutors; in her classroom, with her playing many of the roles that the tutors were constructed to play, demand for feedback from the tutors may have been lower. Anecdotally, this trend was especially present during the first class, wherein Teacher A actively told teams to ignore the tutorial provided by the Mentor and instead follow along with her tutorial on the board.

Despite this difficulty leaving room for the tutors' feedback in the classroom, Teacher A did comment that as the intervention moved along, she found herself depending more on the tutoring system and struggling to avoid mentioning it in classes that did not have access to it:

> *"[In Control group classes], I kept biting my tongue and wanting to say, 'What does the Critic say?'" –Teacher A*

When asked for feedback on how the tutors might be improved, the educators commented on four specific areas for improvement. First, the instructional assistant commented on wanting greater specificity from the tutors' feedback. She remarked that at times, teams became frustrated because the tutors repeatedly provided feedback that they did not know how to incorporate. However, she also comments that this presents a delicate balance because many teams are also discouraged when asked to do a lot of reading:

> *"It depends on the student. From my point of view with my students, more is better at certain times to clarify things. For a lot of them, once they get over the hump, they'll be fine if you give them a little more." –Instructional Assistant*

She also commented that she witnessed on some occasions that the tutors did operate in precisely this fashion, remarking on students' prior improvements before providing additional feedback:

> *"I did see some tutors, especially the Critic, that said 'I like that you did this, but this needs to be worked on', I did see that but maybe just a little more detail." –Instructional Assistant*

Teacher A specifically suggested that the tutors themselves initiate more interaction with the teams by actively interrupting teams rather than passively alerting teams when feedback is available or waiting for teams to initiate interaction. This feedback actually ran directly counter to the design of the system, which was intended to limit the degree to which teams are interrupted based on past research indicating that frequent interruptions from autonomous software agents are often a source of irritation or frustration for user (Schiaffino & Amandi 2004; Rudman & Zajicek 2006). Teacher B commented that she liked the Interviewer's prompts that asked students to write down responses in natural language, even though those responses could not be used for live feedback:

> *"I liked the parts where they had to write, so where they popped up the box and said explain why you did this or what would happen here. Any kind of writing components, I would expand, especially with the new standards coming out in science, writing is a big aspect in science. They have to get that lingo down, so more writing." –Teacher B*

Finally, both Teacher A and Teacher B commented that they believed the teams found the facial expressions of the tutors distracting at times; both suggested that sometimes, teams were more impacted by the facial expression of the tutor than the feedback itself. In some instances, this appeared distracting to the teachers, as teams appeared to interpret positive feedback as negative or negative feedback as positive based on a misinterpretation of the tutor's expression.

Feedback on the Intervention as a Whole

Finally, educators were asked to reflect on the intervention and study as a whole, including phenomena that teams investigated, the classroom activities outside the software, and the learning objectives of the unit. Both Teacher A and Teacher B appreciated the classroom activities that went beyond the software, such as the model walks and the lake water exercise, and specifically liked how those activities tied back into the exercise within MILA. Teacher A noted some specific positive student feedback:

> *"The students loved this, I got comments like 'I've never done anything like this before in science. This is awesome.' They really enjoyed it. They still want to know what happened to the fish." – Teacher A*

Teacher A also suggested elongating the unit to allow for an additional "training" project, adding in additional hands-on activities to provide data for explanations in

MILA, and supplying additional readings on the phenomena. Teacher B similarly suggested expanding the number of activities outside the software that could provide data for explanations in MILA, and either limiting the number of "model walks" reviewing others' models or providing clearer differences in objective to the existing model walks.

Takeaways from Educator Feedback

There are a number of takeaways from the results of the interviews with the instructors that help contextualize the results in the prior analyses. First, the feedback from the teachers helps to establish an understanding of the way in which the classroom intervention manifested at the classroom level rather than at the team level. Earlier analysis of the results of the modeling exercise showed that teams in the Experimental group outperformed teams in the Control group on several evaluations of model quality. However, in some areas, these improvements were difficult to trace back to the tutoring system; in fact, some of these improvements came in areas that the tutoring system was not equipped to address. Based on that analysis, there was speculation that the benefits could have come from the way the presence of the tutoring system changed the classroom as a whole, handling easier questions and feedback opportunities and allowing the teachers to focus on deeper feedback and trends at the classroom level. Feedback obtained from the teachers appears to corroborate this idea: the teachers commented that they were aware of the tutoring system playing some of the roles they traditionally played, and more importantly aware that they were not having to answer as many questions and provide as much feedback in the Experimental classes. This interplay between the individual teams using the tutors and the patterns of interaction amongst the classroom as a whole may be the most significant trend to originate from this study. This possible explanation of the results described in Chapter 4 is discussed in further detail in Chapter 6, under the section "Discussion of Primary Claim" on page 216.

Second, a related takeaway is that the benefits of access to MILA–T may have been partially felt not by the students, but rather by the teachers. With MILA–T taking care of some of the simpler questions about model quality and completeness, software interaction, and lesson goals, educators may have been more free to focus on more advanced commentary or trends that spanned multiple teams and classes instead of spending their time answering repetitive questions. Of course, in order for this to be a benefit, the effect of this dynamic would ultimately have to trickle down to the individual students in different ways, such as students receiving better feedback than they would have otherwise received. Thus, the effects of the dynamic should still be captured in some way in the data analyzed above. Additional analysis in future studies, however, ought to explicitly analyze the frequency, nature, and complexity of questions that students ask of the teacher and feedback that the teacher provides to the students. The observations made here suggest that teachers supported by tutoring systems may be able to provide better feedback to students than teachers forced to provide all the feedback on their own.

A third takeaway builds on the design of the individual tutoring agents. As described in chapter 2, each of the five agents is specifically constructed to mimic a functional role that the teacher plays in the classroom. These are functional roles that the teacher is accustomed to playing, based on Grasha's model of teaching styles. The nature of these functional roles leads to two possible results: either the teacher and the tutoring system clash over what advice to give (as seen in Teacher A's initial experience), or the tutoring system allows the teacher to release certain functional roles and focus more on others (as reflected more in the feedback from Teacher B and the instructional assistant). This suggests that implementing a tutoring system in a classroom environment does require a learning curve on the part of the teacher; many habits must be temporarily unlearned in order to take full advantage of the support provided by the tutoring system.

CHAPTER 6

CLAIMS

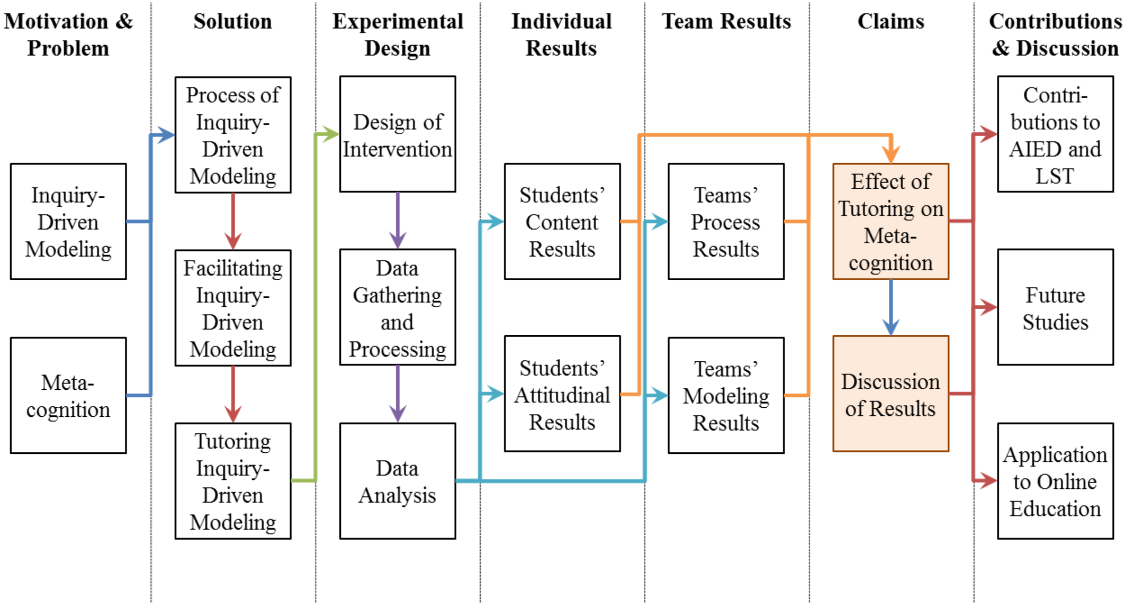

Motivation & Problem
Solution
Experimental Design
Individual Results
Team Results
Claims
Contributions & Discussion
Inquiry-Driven Modeling
Meta-cognition
Process of Inquiry-Driven Modeling
Facilitating Inquiry-Driven Modeling
Tutoring Inquiry-Driven Modeling
Design of Intervention
Data Gathering and Processing
Data Analysis
Students' Content Results
Students' Attitudinal Results
Teams' Process Results
Teams' Modeling Results
Effect of Tutoring on Meta-cognition
Discussion of Results
Contri-butions to AIED and LST
Future Studies
Application to Online Education

Figure 27: The sixth chapter of this dissertation summarizes the effect of the metacognitive tutoring system identified in chapters 4 and 5. It then discusses three different explanations of how the observed effect may have arisen.

This sixth chapter will synthesize the results of the prior analysis into a set of final claims that will set up further explanation of this work's contributions in Chapter 7. Chapter 6 will start by recapping the individual claims made as a result of the analysis above on each of the four hypotheses. It will briefly summarize the analyses and results of each of the prior for hypotheses and restate the ultimate claim (or lack thereof) associated with each. It will then synthesize those individual hypotheses into a broader claim about the extent to which this intervention actually improved metacognition in the students using the metacognitive tutoring system. This chapter begins by revisiting the original hypotheses, shown in Table 87 below.

Table 87: The eight hypotheses of this study.

Explicit Understanding	
Hypothesis #1: Engagement with a metacognitive tutoring system during participation in an inquiry-driven modeling task will improve students' declarative understanding of inquiry-driven modeling compared to participation in inquiry-driven modeling without a metacognitive tutoring system.	
Dispositional Framing	
Hypothesis #2: Engagement with a metacognitive tutoring system during participation in an inquiry-driven modeling task will improve students' dispositional framing of science, scientific inquiry, and careers in science compared to participation in inquiry-driven modeling without a metacognitive tutoring system.	
Procedural Execution	
Hypothesis #3: Engagement with a metacognitive tutoring system during participation in an inquiry-driven modeling task will improve teams' execution of the desired inquiry-driven modeling process compared to participation in inquiry-driven modeling without a metacognitive tutoring system.	
Hypothesis #3A: This improved execution of the desired inquiry-driven modeling process will take place while the team is receiving feedback from the metacognitive tutoring system.	**Hypothesis #3B:** This improved execution of the desired inquiry-driven modeling process will take place when the team is no longer receiving feedback from the metacognitive tutoring system.
Models and Explanations	
Hypothesis #4: Engagement with a metacognitive tutoring system during participation in an inquiry-driven modeling task will improve the quality of models and explanations that teams generate compared to participation in inquiry-driven modeling without a metacognitive tutoring system.	
Hypothesis #4A: These improvements will be present in modeling activities during which the team is receiving feedback from the metacognitive tutoring system.	**Hypothesis #4B:** These improvements will be present when the team is no longer receiving feedback from the metacognitive tutoring system.

This study began with the overall hypothesis that engagement with a metacognitive tutoring system (MILA–T) during an inquiry and modeling activity in an exploratory learning environment (MILA) would lead to improved metacognition of inquiry-driven modeling in participants compared to engagement in an inquiry and modeling activity in an exploratory learning environment without a metacognitive tutoring system. This broad hypothesis was further decomposed into four sub-hypotheses regarding explicit understanding, dispositional framing, the modeling process, and models and explanations, each of which represented a portion of the metacognitive ability of inquiry-driven modeling as derived from existing literature on the components of a metacognitive skill (Roll et al. 2007). These hypotheses are duplicated in Table 87 above.

Hypothesis #1: Explicit Understanding

The first hypothesis of this study is that engagement with a metacognitive tutoring system during participation in an inquiry-driven modeling task will improve students' declarative understanding of inquiry-driven modeling compared to participation in an inquiry-driven modeling task without a metacognitive tutoring system. In order to assess this, students were given a content test at the beginning and end of the intervention; this content test is supplied in Appendix B. The content test was developed during the spring of 2012 by an expert in education and educational assessment, but it has not been independently validated. The content test emphasizes metacognitive knowledge of scientific inquiry and modeling, as well as certain metacognitive elements of biological content knowledge (such as understanding the role of systems and the interaction between invisible and visible phenomena), as best as can be established through a multiple-choice test.

Prior to the intervention, these existed no difference between the Control and Experimental groups in performance on the content test. After the intervention, there still existed no difference between the Control and Experimental groups in performance on the content test. Students in both groups exhibited statistically significant improvement in their content test scores, but that improvement was not tied to Condition. Therefore, there are insufficient data to accept that engagement with a metacognitive tutoring system during participation in an inquiry-driven modeling task improved students' declarative understanding of inquiry-driven modeling compared to participation in an inquiry-driven modeling task without a metacognitive tutoring system. Thus, **Hypothesis #1 is not accepted**. The full analysis of Hypothesis #1 can be found beginning on page 100.

Hypothesis #2: Dispositional Framing

The second hypothesis of this study is that engagement with a metacognitive tutoring system during participation in an inquiry-driven modeling task will improve students' dispositional framing of science, scientific inquiry, and careers in science compared to participation in an inquiry-driven modeling task without a metacognitive tutoring system. In order to assess this, students were given an attitudinal survey at the beginning and end of the intervention. The items on the survey measured five constructs: Attitude toward Scientific Inquiry, Career Interest in Science, Anxiety toward Science, Desire to Do Science, and Perception of the Science Teacher. The first two originate from the Test of Science-Related Attitudes (TOSRA; Fraser 1981), and the last three originate from the modified Attitudes toward Science Inventory (Weinburgh & Steele 2000). Together, these inventories measure students' overall attitude toward science as well as their specific interest in scientific inquiry and the real-life process of science.

Prior to the intervention, these existed no significant difference based on Condition between the Control and Experimental groups. After the intervention, there existed significant differences between the Control and Experimental groups. Students in the Experimental group had statistically significantly higher scores on the Attitude toward Scientific Inquiry construct than students in the Control group. Additionally, there existed a statistically significant interaction between the Condition and the change in

Attitude toward Scientific Inquiry. Over the course of the intervention, the Experimental group also exhibited a statistically significant improvement in their Career Interest in Science that the Control group did not exhibit. No such interactions were observed in Anxiety toward Science, Desire to Do Science, or Perception of the Science Teacher, although students in both the Control and Experimental groups did exhibit a significant drop in Anxiety toward Science. Based on the improvements in the Experimental group to students' Attitudes toward Scientific Inquiry and Career Interest in Science, we can conclude that engagement with a metacognitive tutoring system during participation in an inquiry-driven modeling task improved students' dispositional framing of science, scientific inquiry, and careers in science compared to participation in an inquiry-driven modeling task without a metacognitive tutoring system. Thus, **Hypothesis #2 is accepted**. The full analysis of Hypothesis #2 can be found beginning on page 103.

Hypothesis #3: Procedural Execution

The third hypothesis of this study is that engagement with a metacognitive tutoring system during participation in an inquiry-driven modeling task will improve teams' execution of the desired inquiry-driven modeling process compared to participation in an inquiry-driven modeling task without a metacognitive tutoring system. This hypothesis applies in two different dimensions. First, this improved execution of the desired inquiry-driven modeling process will take place while the team is receiving feedback from the metacognitive tutoring system, and second, this improved execution of the desired inquiry-driven modeling process will take place when the team is no longer receiving feedback from the metacognitive tutoring system. Put differently, teams will **learn** an improved process of inquiry-driven modeling as measured by execution during the receipt of feedback, and teams will also **transfer** this improved process to a new task where no such extra feedback is present.

To measure teams' execution of inquiry-driven modeling, detailed logs of software interaction were kept throughout the intervention. After the intervention, these logs were mapped into Markov chains representing the patterns of interaction in which teams engaged during the two projects. These Markov chains, in turn, were connected to the original model of inquiry-driven modeling, found in Figure 4 on page 25, that motivated the design of the metacognitive tutoring system in order to measure improved execution of the desired underlying process. Nine individual desirable effects were identified after designing, but prior to compiling the data for, the Markov chains describing the activity of the Control and Experimental groups; these desirable effects are shown in Table 44 on page 138. After compiling the Markov chains, the Control and Experimental groups were tested for differences in each of these nine identified desirable effects within the Learning project and Transfer project independently. In both the Learning and the Transfer project, differences were observed in three of the nine tests. In each project, the Experimental group performed better in two of the tests while the Control group performed better in one of the tests. Due to the presence of effects supporting both groups, and due to the increased likelihood of false positives based on the repeated Z-test, no conclusions can be drawn based on these data. A subsequent χ^2 test demonstrated more thoroughly that teams in the Control and Experimental groups exhibited different patterns of interaction during each project, but this test does not provide any conclusions about the superiority of one group's pattern of interaction over the other. Additional analysis provided more detail on the way in which teams interacted with the tutoring system and the way in which those interactions altered the overall pattern of inquiry and modeling; however, the comparative analyses of the Control and Experimental groups' processes did not provide sufficient evidence to conclude either group performed conclusively better than the other. Therefore, there are insufficient data to conclude that engagement with a metacognitive tutoring system during participation in an inquiry-driven modeling task improved teams' execution of the desired inquiry-driven modeling process compared to participation in an inquiry-driven modeling task without a metacognitive tutoring system, either during initial interaction with the tutoring system or during subsequent activity without the tutoring system. Therefore, **neither Hypothesis #3A nor Hypothesis #3B are accepted**, and thus, **Hypothesis #3 is not accepted**. The full analysis of Hypothesis #3 can be found beginning on page 124.

Hypothesis #4: Models and Explanations

The fourth hypothesis of this study is that engagement with a metacognitive tutoring system during participation in an inquiry-driven modeling task will improve the quality of models and explanations that teams generate compared to participation in an inquiry-driven modeling task without a metacognitive tutoring system. As with the third hypothesis, this hypothesis applies in two different dimensions. First, these improvements will be present in modeling activities during which the team is receiving feedback from the metacognitive tutoring system, and second, these improvements will be present when the team is no longer receiving feedback from the metacognitive tutoring system. Put differently, teams with the metacognitive tutoring system will put together better models and explanations both while receiving feedback on their models and after they no longer are receiving feedback on their models.

To measure this, five different metrics were designed to summarize the quality of teams' models, both in terms of the model's complexity and in terms of the strength of the evidence that teams supply in defense of their model. The finals models submitted by teams at the conclusion of both the Learning and the Transfer project were processed and summarized according to these metrics. Then, models were compared within each project. In both the Learning and the Transfer project, teams in the Experimental group generated models that rated more highly on three of the five metrics for model evaluation. However, these metrics for model evaluation relied on teams accurately categorizing the evidence they supplied in defense of their explanations. This desired accurate categorization involved both correctly identifying the category for good pieces of evidence as well as providing good pieces of evidence in the first place. In order to assure this, every connection was coded for the acceptability of its evidence. At the conclusion of this process, a χ^2 was run on the results of the coding. This test revealed that teams in the Experimental group performed better on their evidential justifications than teams in the Control group in both the Learning and Transfer projects. Both groups had equal rates of Miscategorized evidence, but a significantly lower proportion of evidence supplied by teams in the Experimental group was considered unacceptable compared to the Control group. Then, using the results of this coding exercise, the four evidence-based metrics were recalculated for each model and the original analysis was performed again. This time, models produced by teams in the Experimental group were rated as even better compared to models produced by teams in the Control group than in the previous analysis. In the Learning project, teams in the Experimental group generated better models according to three metrics, while in the Transfer project, teams in the Experimental group generated better models according to all five metrics. Each of these three analyses provides evidence that shows teams in the Experimental condition produced better-justified models than teams in the Control condition. Therefore, there are sufficient data to conclude that engagement with a metacognitive tutoring system during participation in an inquiry-driven modeling task improved the quality of the justifications that teams provided for their models and explanations compared to participation in an inquiry-driven modeling task without a metacognitive tutoring system, and that the improved quality of justifications for their models applied both while receiving feedback from the tutoring system and after feedback from the tutoring system had been disabled. Therefore, **both Hypothesis #4A and Hypothesis #4B are accepted**, and thus, **Hypothesis #4 is accepted**. The full analysis of Hypothesis #4 can be found beginning on page 151.

Overall Claims

The analysis shows that there exist data to accept three of the original six hypotheses. There do not exist data to support the remaining three hypotheses. Table 88 below summarizes the conclusions of this analysis; hypotheses shaded in green are accepted, while hypotheses shaded in red and yellow are not. Red shading indicates that there are no data to support the hypothesis, while yellow shading indicates that the data are mixed with regard to whether or not they support the hypothesis.

Table 88: Summary of the three accepted hypotheses of this study (shaded in green) and the three rejected hypotheses (shaded in red or yellow).

Hypothesis #1: Explicit Understanding	
Hypothesis #2: Dispositional Framing	
Hypothesis #3A: Procedural Execution (Learning)	**Hypothesis #3B: Procedural Execution (Transfer)**
Hypothesis #4A: Models & Explanations (Learning)	**Hypothesis #4B: Models & Explanations (Transfer)**

Clearly, these results are mixed. The literature defines metacognitive skills as consisting of at least three parts: declarative, procedural, and dispositional understanding (Roll et al. 2007). This analysis found no evidence for improved declarative understanding of inquiry-driven modeling as a result of engagement with the metacognitive tutoring system. This analysis also found no conclusive evidence for improved procedural understanding of inquiry-driven modeling as a result of engagement with the metacognitive tutoring system. However, this analysis did find that the results of the inquiry-driven modeling process were superior among teams who had received the metacognitive tutoring system compared to those who had not, and this improvement applied both while the teams were receiving feedback from the metacognitive tutoring system and after the metacognitive tutoring system had been disabled. Given that these explanations were the direct output of the inquiry-driven modeling process in which groups engaged, it stands to reason that there did exist improvement in the inquiry-driven modeling process that was not captured by the data-gathering methods present in this study. Finally, this analysis also showed that students' dispositional understanding of science, scientific inquiry, and careers in science improved based on engagement with the metacognitive tutoring system.

The absence of evidence that students' declarative understanding improved and the absence of definitive evidence that students' procedural execution improved, however, makes it difficult to conclusively accept the original hypothesis. Therefore, this analysis does **not** claim to have shown that engagement with a metacognitive tutoring system (MILA–T) during an inquiry and modeling activity in an exploratory learning environment (MILA) leads to improved metacognition of inquiry-driven modeling in participants compared to engagement in an inquiry and modeling activity in an exploratory learning environment without a metacognitive tutoring system. There are insufficient data to conclude that metacognition itself has been improved, and thus, there are insufficient data to accept the original hypothesis of this study.

While there are insufficient data to accept the original hypothesis of the study, there exist sufficient data to conclude that engagement with the metacognitive tutoring system nonetheless had strong positive effects on students' learning. Students who participated in the modeling and inquiry tasks with the help of the metacognitive tutoring system showed a strong increase in their attitudes toward scientific inquiry and their career interest in science as compared to students that participated in the same process without the metacognitive tutoring system. Teams who had access to the metacognitive tutoring system produced better-justified explanations of the target phenomenon than teams that did not, and more importantly, they continued to produce stronger explanations of the phenomena after the tutoring system had been disabled. Thus, the lessons learned from the tutoring system during the initial phase of engagement transferred to a new task. Therefore, these improvements cannot be explained as simply reactions to increased feedback, but rather may be described as an internalized understanding of the criteria of a good explanation. While it is difficult to make strong claims about the students' metacognition itself based on these results, these results are sufficient to conclude that engagement with the metacognitive tutoring system was beneficial to students' attitudes toward science and development of explanations of scientific phenomena.

Thus, the ultimate claim of this analysis is: **engagement with a metacognitive tutoring system during participation in an inquiry-driven modeling task improves students' attitudes toward scientific inquiry and careers in science, and also improves students' ability to generate well-justified models and explanations of a phenomenon both during and after engagement with the tutoring system**.

Discussion of Primary Claim

The above claim is made based on the data present in the analysis; there are sufficient data to accept the hypotheses that participation in the activity with the metacognitive tutoring system improved attitudes toward scientific inquiry, career interest in science, the quality of models that teams produced, and the quality of evidence that teams produced. However, this analysis does not independently describe the reason why these changes actually took place. For that, there exist multiple possible explanations for the effect of the metacognitive tutoring system on the observed variables. Three alternatives are discussed here: authentic engagement, simple reinforcement, and classroom-level interactions.

Authentic Engagement

The first, and arguably simplest, explanation for the observed effects is that MILA–T worked the way in which it was intended to work. As teams interacted with MILA, they received feedback from the tutoring system on the progress they were making and areas for improvement. As they received this feedback, they applied it and improved their explanations accordingly. As a result, their models and evidence during the Learning project were improved. The advice that the teams received from the tutoring system was not simply applied to the current models; the teams instead internalized the advice, thus transferring it to the Transfer project. When teams were confronted with a new phenomenon to model, they remembered the suggestions and advice of the tutors during the Learning project and continued constructing superior models because of the feedback they had previously received. As a result, their models and evidence during the Transfer project were improved. Because of the feedback received from the tutoring system, the teams were also more able to completely engage in the authentic process of inquiry and modeling, receiving a better view of the real process of science than groups relying only on teacher feedback for improvement. As a result of this exposure to and participation in authentic science, Experimental students also improved in their attitudes toward inquiry and career interest in science because of the inherent appeal of these elements when presented authentically.

This explanation provides a somewhat straightforward explanation for how access to this tutoring system led to the observed differences between Control and Experimental groups; however, it also makes a potentially flawed assumption. As noted in the results, one of the advantages demonstrated by the Experimental teams was superior evidence; an average of 13% more of the evidence written by Experimental teams was deemed acceptable. The tutors of MILA–T, however, are unable to read the evidence written by students. Under the authentic engagement explanation of these results, the improvement seen in the Experimental teams relies on the assumption that teams learned to write better evidence simply from the direct instruction of the Guide rather than from any other source. While possible, it is perhaps unlikely that such a significant improvement could be seen simply from access to a couple more paragraphs of text that many groups did not even access. For those data, the second or third explanations may provide more compelling accounts.

Simple Reinforcement

In order to describe the way in which access to MILA–T led to the gains documented previously, it is important to consider the role of superficial elements of the presence of the tutoring system in addition to the actual feedback it provides. In doing so, a somewhat cynical explanation of the results observed above may emerge.

As teams in the Experimental group interacted with MILA, they received reinforcement from the tutoring system on the progress that they made. The Mentor would show up to praise their improvement, and the Critic would give very clear indicators of progress as teams expanded their model. The Critic would also give teams clear descriptions of areas for improvement, providing a clear goal toward which Experimental teams could strive. As a result, their models during the Learning project were improved. Teams in the Control group, on the other hand, received no such consistent and reliable reinforcement, and only received any feedback at all after waiting for the teacher to come around and provide feedback. Control teams, thus, also had limited or intermittent access to an external agent setting goals for them, as well as intermittent access to suggestions on areas for improvement. In this way, the modeling advantages seen by the Experimental group could be descried as simple responses to reinforcement. The presence of reinforcement improved students' attitudes toward the activity itself; they received more consistent indicators of progress and success than Control teams, and therefore improved in their attitude toward the task at hand. As a result of their positive perception of the activity, Experimental students also improved in their attitudes toward inquiry and career interest in science.

This explanation runs into difficulty with two of the data points, however. First, one of the observed advantages of the Experimental teams was that they wrote superior evidence. The tutors of MILA–T do not give any feedback on the actual evidence that teams write, however; it only responds to the categories. If simple reinforcement is the explanation of the results, then this explanation encounters difficulty explaining an improvement that lacked direct reinforcement. A different kind of reinforcement, however, could make up for this gap. Although teams in the Experimental group never received feedback from MILA–T on the content of their written evidential justifications, they may have nonetheless perceived the tutoring system as reading their evidence; the presence of human faces alongside the tutoring system may further enhance this dynamic. This presence of the Hawthorne effect (Cook 1962) may explain this superior performance. An alternative explanation along this same line of thought is that the presence of a human face alongside the tutors humanizes them, leading to a perception among the students that their work is valued and supported; this leads to increased authentic engagement and increased effort into writing good evidential justifications. This echoes the literature's past support for intelligent tutoring systems with human-like names and faces (e.g. Lepper & Chabay 1988; Johnson, Rickel & Lester 2000; Azevedo et al. 2009; Lee & Ko 2011).

This explanation effectively suggests that the *nature* of the feedback teams receive from the tutoring system is irrelevant to the results seen here; the benefits can be explained simply by the *presence* of feedback, regardless of the actual content of the feedback. This explanation, however, has some difficulties explaining the entirety of the data as well. Specifically, this explanation encounters some difficulty with the models and evidence generated during the Transfer project. As documented previously, the advantages seen in the Experimental group during the Learning project transferred to the Transfer project. However, during the Transfer project, the Experimental teams received no such feedback, limiting the potential presence of the Hawthorne effect. It is possible that the persistent advantages seen by the Experimental teams here were due solely to the positive habits cultivated through the availability of feedback during the Learning project, but this explanation is weaker than the prior explanation for the positive results in the Learning project and in attitudes overall.

Classroom-Level Interactions

The above two explanations describe two ways in which the presence of feedback could lead to the advantages documented previously: either the specific nature of the feedback led to these improvements, or the existence of feedback altogether led to these improvements. Both explanations encountered difficulty explaining the improved evidence that teams in the Experimental condition wrote because neither explanation can provide a particularly compelling account for where Experimental teams would receive feedback on their evidential justifications. The authentic engagement explanation relies on the effectiveness of rote instruction while the simple reinforcement explanation demands the presence of a sort of Hawthorne effect, and while these are possible, the advantages seen are perhaps too great to explain without some sort of additional feedback and resultant improvement in the Experimental teams.

If the Experimental teams did not receive feedback from the tutoring system on the text of their evidential justifications, but yet their evidential justifications were superior to those teams without the tutoring system, from where could this improvement have arisen? The third explanation is that the presence of the tutoring system interacted with the classroom as a whole to lead to these gains. As explained in Chapter 3 during the description of the curriculum, teachers in the classes were also responsible for giving feedback to students during the Learning and Transfer projects. Unlike the tutors of MILA–T, however, the teachers were capable of actually providing feedback on the text that teams wrote. Therefore, the superior evidential justifications in the Experimental teams could be described as a response to feedback from the teachers. This introduces a second question: why would the teams in the Experimental condition receive better feedback from the teachers than teams in the Control condition? To explain this, we can refer back to the teachers' own reports on their behavior in the classroom described in the final section of Chapter 5. The teachers referenced their reliance on the tutoring system to give feedback to multiple groups at the same time, therefore lessening the demand on the teachers. This may lead to multiple effects. First, because the teachers are spending less time giving individual groups feedback, they are better able to focus on classroom-level trends outside the scope of what MILA–T can observe. Based on these trends, they can address broader misconceptions amongst multiple groups, improving the exercise as a whole for all teams in the classroom. Second, because students are able to receive instantaneous feedback from the tutoring system during their model construction process, they are able to ascend to the level of needing teacher feedback more quickly. The teachers can then provide feedback to those groups who have already exhausted the capabilities of the tutoring system and are thus ready for more advanced feedback, such as suggestions on improving the actual nature of their evidence.

In this way, the effectiveness of MILA–T in improving the models and evidence generated by Experimental teams may be described as the successful offloading of functional roles from the teacher to the tutoring system. By taking care of some of the teacher's responsibilities, the tutoring system allows the teacher to focus on those roles that only he or she is still equipped to play. Teams are able to iterate over their explanations quickly with MILA–T's immediate feedback, meaning that they can rise to the level of demanding the kind of feedback only the teacher can provide more quickly. Then, the teacher can supply advanced feedback to a greater number of groups in the same time period. As a result, their models and evidence during the Learning project improved. Like the authentic engagement explanation and unlike the simple reinforcement explanation, this explanation suggests that the effectiveness of the tutoring system is derived from the actual feedback that is being provided. As teams receive this feedback, they internalize it rather than simply respond to it. As a result, their models and evidence during the Transfer project improved. Enhancing this trend even more, these Experimental teams receive and internalize feedback not only from the tutoring system, but also from the teachers. As with the authentic engagement explanation, this process led to students more deeply engaging in the authentic process of inquiry, and as they are exposed to the true nature of inquiry and careers in science, their attitudes improved. The presence of classroom-level interactions may also explain even broader lessons outside the software that the teacher was able to communicate as the demands on the teacher's time lessened, further explaining the attitudinal gains seen in this study.

Summary of Discussions

This section has presented three competing explanations for how access to MILA–T may have led to the observed improvements. In actuality, it is unlikely that only one of these explanations was at play. In all likelihood, the observed advantages of the Experimental students and teams are based on a combination of both these explanations and other unidentified factors. The authentic engagement explanation almost certainly plays a role in generating the improvements documented previously as engagement with the tutoring system identifies specific, explicit areas where students can improve their current explanation. While the simple reinforcement explanation relies on these specific identified areas as well, it also relies on some authentic engagement to explain how students know *how* to improve. The classroom-level interactions explanation, similarly, relies on some authentic engagement to accelerate teams to the point of even being prepared for the type of feedback that led to the advanced improvements documented previously. It would also be disingenuous to discount the effect of the simple reinforcement explanation while accepting the value of the other explanations. Even as the content of the feedback and the presence of the teacher make a difference to student performance, the presence of clear indicators of progress gives students a goal to pursue even if they do not yet understand the value or nature of the goal. The presence and influence of human faces dynamically responding to students' actions should also not be discounted. But while this explanation may help explain why students want to do better, it similarly cannot explain how they know *how* to do better. While the authentic engagement explanation would point out that the tutoring system provides information on the role of evidence, it relies on the assumption that students will internalize that information without receiving individualized feedback. The classroom-level interactions explanation provides a more compelling explanation as teams receive feedback on how to do better from an agent acting outside the purview of the software: the teacher.

Similarly, it is almost certain that portions of the effects seen in this study can and should be explained by mechanisms outside of these three. For example, one could take an apprenticeship approach (similar to Lave & Wenger 1991) to explaining these effects and suggest that the superior performance seen initially was because the tutoring agents were doing some of the work for the Experimental teams; in the Transfer project, the teams had more experience with some of the skills and were thus able to focus on the portions of the work that had previously been completed by the tutoring system. One could also argue that the information and feedback provided by the tutoring system amounted to only minor improvements, but those improvements in this study facilitated a tipping point that cascaded to greater improvements; the improvements derived from the tutoring system were simply located in critical places.

Several other explanations may also be offered, and the continued experimentation outlined in Chapter 7 aims to help narrow down the optimal explanation for the observed effects. Specifically, it is important to isolate the multiple competing variables at work in this study: the nature of the feedback; the level of the feedback; the presence of human faces on the tutoring system; the times at which the tutoring system interrupts; the rate at which teams may solicit feedback from the tutoring system; the presence of the human teacher; the roles played by the human teacher; the roles played by the tutoring system; the curriculum accompanying the tutoring system; the exploratory learning environment itself; and the context in which the exploratory learning environment and intelligent tutoring system are used. Analysis of these variables will help identify the most compelling explanation for the observed results.

CHAPTER 7

CONTRIBUTIONS AND DISCUSSION

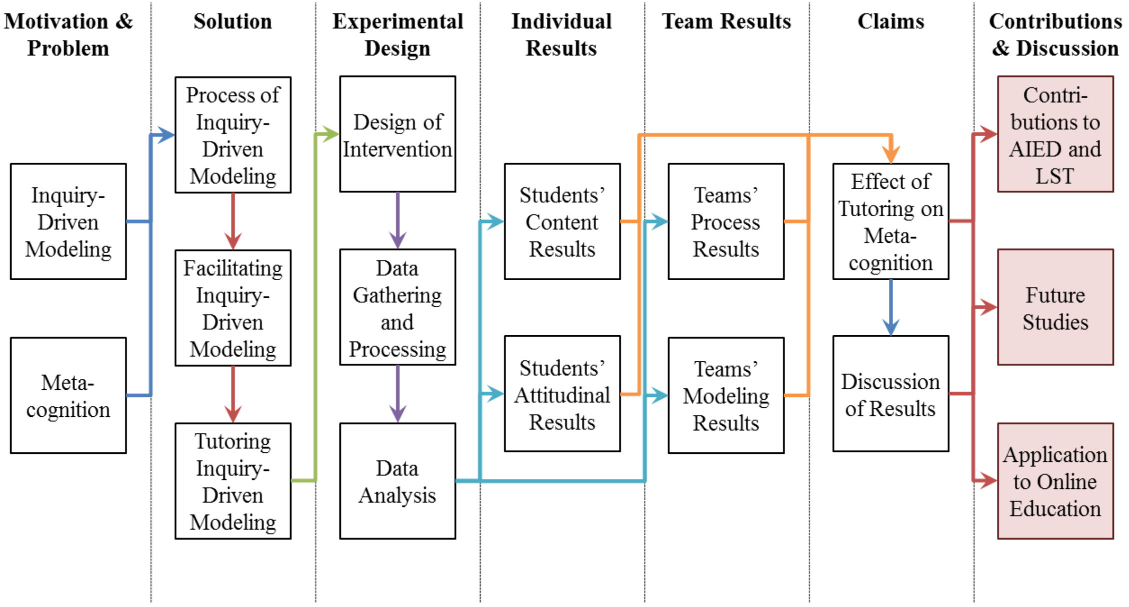

Motivation & Problem
Solution
Experimental Design
Individual Results
Team Results
Claims
Contributions & Discussion
Inquiry-Driven Modeling
Meta-cognition
Process of Inquiry-Driven Modeling
Facilitating Inquiry-Driven Modeling
Tutoring Inquiry-Driven Modeling
Design of Intervention
Data Gathering and Processing
Data Analysis
Students' Content Results
Students' Attitudinal Results
Teams' Process Results
Teams' Modeling Results
Effect of Tutoring on Meta-cognition
Discussion of Results
Contri-butions to AIED and LST
Future Studies
Application to Online Education

Figure 28: The seventh chapter of this dissertation covers the contributions of this work to two communities, artificial intelligence in education and learning sciences & technology. It then covers future studies that ought to be conducted to expand on the findings here, as well as the way in which the principles at work in this dissertation have been applied in a different domain.

The previous chapter described the primary claim of this work: engagement with a metacognitive tutoring system during participation in an inquiry-driven modeling task improves students' attitudes toward scientific inquiry and careers in science, and also improves students' ability to generate good models and explanations of a phenomenon both during and after engagement with the tutoring system. Building on that primary claim, Chapter 7 will list the specific individual contributions of this work to a variety of communities, including the artificial intelligence for education community, the learning sciences and technology community, and the human-computer interaction in education community. Finally, this chapter will discuss opportunities for future research building from the work described in this study, as well as recent applications of the same high-level goals and design principles in new settings.

Contributions

The primary contribution of this work is the claim stated above: the presence of a metacognitive tutoring system during participation in an inquiry-driven modeling task can improve students' attitudes towards scientific inquiry, interest in science careers, and explanatory models of observable phenomena. A variety of mechanisms may be at play to create this change, but the ultimate contribution is the statement that the presence of such a system can lead to such valuable gains.

In addition, however, a number of other contributions can be derived from this work. We anticipate major contributions to two communities and fields of research: artificial intelligence in education and learning sciences & technology. These contributions will likely also connect to other communities as well, such as intelligent tutoring systems, human-computer interaction in education, scientific modeling in education, exploratory learning environments, and inquiry-based learning.

To Artificial Intelligence for Education

The primary community for the work presented here is the artificial intelligence for education community given that the artificially intelligent tutors of MILA–T are the main target of this study. A number of contributions can be derived from the design, development, and deployment of this tutoring system, including the existence of the tutors themselves, a new design idea for choosing the specific roles that these tutors will play, and a theory regarding the impact of these tutors on a classroom setting.

MILA–T

The first and most obvious contribution of this work is the tutoring system itself. As a result of this project, there exists a set of five tutors which together help students improve their understanding of inquiry-driven modeling. Based on the results of the analysis presented here, the usefulness of this tutoring system is documented:

engagement with an activity using this tutoring system led to improved attitudes toward science, interest in science as a career, and explanations of the target phenomenon both while and after feedback from the tutoring system was available.

The tutors of MILA–T are a contribution to the artificial intelligence in education community in two ways. First, in conjunction with MILA, both are available for reuse and deployment. While the tutors of MILA–T are designed with small affordances made to the specific curriculum used during the intervention described here, their overall operation is general over any long-term modeling and inquiry project. While the tutors of MILA–T are broadly constructed to operate in the domain of ecology, within this area they are general to the individual phenomenon; no changes need be made to apply MILA–T to investigations of ecological phenomena besides fish kills (although additional content knowledge may be valuable) so long as the tutors are operating with CMP models. This applies to MILA as well; although the specific syntax of the CMP models used in MILA restricts the interface to modeling ecological phenomena, within this general field any new phenomenon can be approaching, modeled, and investigated within the system.

Secondly, beyond engagement with MILA and MILA–T as presently constructed, the design of the tutoring system is also generalizable out to other modeling domains as well. The development of the Guide, the Critic, the Mentor, the Interviewer, and the Observer in other domains and with other modeling frameworks would require specific connections to the semantics and structure of the different representations, but those connections are largely compartmentalized away from the higher-level design of the tutors. So long as the information used by the individual tutor exists (such as a log of actions for the Interviewer, a presently readable model of the phenomenon for the Guide, or a set of criteria for a good answer for the Critic), the overall pattern of reasoning of the tutor can be abstracted and transferred to new software settings. This has already been partially accomplished as well; in the "Application to MOOCs" section of this chapter on page 245, we describe the appropriation and usage of the design patterns of the Critic in an online course in a radically different domain to provide immediate, personalized feedback on multiple types of problems.

Paradigm for Constructing Functional Roles of Tutors

In addition to interacting with the tutors themselves and appropriating the specific rules that govern their interaction in a modeling task, the MILA–T tutors contribute an additional design principle to the design of cognitive and metacognitive tutors. One of the unique elements of the tutors in MILA–T is their construction to mimic the functional roles played by a teacher in the classroom. To date, we are not aware of any other tutoring system that has explicitly structured the *functional roles* of a tutoring system with attention paid to the typical interactions with a teacher in a traditional classroom; other work has examined human tutoring as a source of inspiration for artificially intelligent tutoring agents (Merrill, Reiser, Ranney & Trafton 1992; Heffernan & Koedinger 2002; VanLehn et al. 2003), but without explicit attention to the functional roles of a teacher in a traditional classroom context or the potential interaction between an intelligent tutoring system and an in-person teacher. Instead, in many tutoring systems, the tutors implicitly take on a narrow understanding of the teacher's role in providing correction and feedback on a present answer. This process is valuable and provides the chief motivating role for the design of the Critic in this work, but it represents a somewhat narrow replication of the range of roles that the teacher plays at any given time. An additional functional role identified in other tutoring systems is the ability to give an appropriately difficult new problem. The majority of tutoring systems build on these two roles.

A close analogue to this idea can be seen in MetaTutor, the tutor for developing self-regulation in students constructing an understanding of scientific material (Azevedo et al. 2011). In MetaTutor, tutors are named with similar roles as the tutors seen here, including a Guide, a Planner, a Strategizer, and a Monitor (Azevedo et al. 2009). These agents are constructed similarly to the agents of MILA–T in that they also each are endowed with a functional role. However, the fundamental difference between the agents of MetaTutor and the agents of MILA–T is the origin of the functional role. In MetaTutor, the agents are constructed with functional roles appropriated from the target skill; Planning, Strategizing, and Monitoring are elements of the self-regulatory process that the system aims to instruct. In MILA–T, the agents are constructed with functional roles appropriated from interaction with an instructor; Critiquing, Mentoring, and Interviewing are not skills that the teacher aims to teach the student, but rather are skills that the teacher uses to teach the student some additional learning goals. To put this differently, the Planner, Strategizer, and Monitor in MetaTutor are all instances of the Mentor role in MILA–T; all attempt to demonstrate and apprentice the student in the target skill for which they are constructed.

On its own, the paradigm used to construct the tutors of MILA–T is unique in the intelligent tutoring systems community. What makes this contribution more significant, however, is that it is deeply tied to one of the hypothesized mechanisms for the improvement seen in this study. As referenced above and described again below, one hypothesis regarding the cause of the improvements based on interaction with the metacognitive tutoring system is that the tutors were able to take care of tasks that the teacher usually completes, thus freeing the teacher up to give deeper feedback. If this is the actual mechanism behind these improvements, then the construction of the tutoring system to specifically mimic the teacher's roles is fundamental to the success of the system. Thus, this suggests that this paradigm of constructing tutors to offload functional roles from the teacher may be valuable in exercises wherein tutors are deployed in a classroom context. This, in turn, may open up the additional gains documented previously specifically because this intervention took place in a somewhat traditional classroom.

Theory on Classroom-Level Effects of Team-Level Interventions

In the past, a large number of experiments and analyses involving intelligent tutoring systems have taken place in controlled settings. Others involved leveraging an existing classroom environment, but removing the teacher from the system entirely and fully controlling the course of the classroom activity. At present, we are not aware of any previous intervention using an intelligent tutoring system that has taken place in a traditional classroom setting wherein a teacher remains present and plays his or her normal functional roles during the exercise. In this study, however, the intelligent tutoring system was injected into a classroom activity wherein the teacher maintained her normal relationship and role in the classroom. Research staff were present for technical support, but interaction with the individual students and teams was very limited.

As thoroughly documented previously, an interesting potential effect of this unique experimental design is that the interactions between the individual teams and the tutoring system may have changed the nature of the interaction between the teams and the teachers, and thus, changed the overall patterns of interaction and feedback for the classroom as a whole. As discussed before, additional research is necessary to identify and quantify this effect. Even in the absence of such additional research, however, this notion in its current form provides three related contributions to the intelligent tutoring system community. First, this research provides a new research question that presently is not being addressed by the research community that might have a significant positive effect. Second, aside from the research question, this research provides some early evidence that deploying tutoring systems in controlled environments rather than natural classrooms may be missing a key benefit of such systems, and thus, to maximize benefit, intelligent tutoring systems ought to be used in traditional classrooms with teacher support as well. Third, in addition to posing the hypothesis that such an effect may be present, this work also contributes a preliminary account of how exactly such an effect may be arising. In this account, a tutoring system specifically constructed to mimic certain functional roles of the teacher may offload from the teacher much of the cognitive load of administering the classroom and observing trends. Under this account, the teacher is able to monitor the classroom at a higher level, detect deeper misconceptions, and provide richer feedback because the tutoring system is addressing the questions that would usually dominate the teacher's time.

To Learning Sciences & Technology

In addition to the contributions to the artificial intelligence in education and intelligent tutoring systems communities, this work also provides significant contributions to the learning sciences & technology community separate and apart from the artificial intelligence elements of the project. These contributions also build on the inquiry in education community, the exploratory learning environments community, and the model-based science education community.

MILA

Exploratory learning environments are common in inquiry-based learning settings. River City (Ketelhut 2007) and Co-Lab (van Joolingen et al. 2005) are two examples of exploratory learning environments in which students can pursue an authentic process of inquiry. However, while these environments facilitate the inquiry side of the process, they pay relatively little explicit attention to the role that scientific modeling can play. Other tools, such as Betty's Brain (Biswas, Leelawong, Schwartz, & Vye 2005), INQPRO (Ting, Zadeh, & Chong 2006), and our own earlier tool ACT (Vattam, Goel, & Rugaber 2011; Vattam et al. 2011A), place a strong emphasis on the development of models. However, this is not typically performed in a naturalistic inquiry-based setting. In ACT, for instance, students learned about a system from a pre-written interactive hypermedia rather than through natural observation of the phenomenon itself (Goel et al. 2010).

With this landscape in mind, MILA provides a significant complement to existing exploratory learning environments that teach and scaffold inquiry-based learning or scientific modeling. It is by no means superior to these systems in that the depth of the modeling or the flexibility of the inquiry taught in MILA is significantly diminished compared to the tools that emphasize one skill or the other, but rather MILA fills in a gap in the teaching of modeling and inquiry in exploratory learning environments in the community. Through the design of MILA, students are able to perform both modeling and inquiry in tandem, using each to guide, structure, and support the other in an authentic replication of naturalistic scientific modeling and inquiry in the real world.

While MILA–T plays an important role in accomplishing this learning goal, the structure of MILA on its own is itself a contribution to the research on exploratory learning environments and inquiry-based learning. Towards the inquiry side, MILA unites multiple elements of the inquiry process into one contained interface. In this study, MILA delivered simulation capabilities for examining computational recreations of the natural phenomena, as well as note-taking capabilities for synthesizing information gathered outside of MILA. In prior studies with MILA as well, additional inquiry-oriented features were included, such as an embedded document browser for searching through data and expert assessments, a pluggable framework for invoking any Flash-based simulation within the program, and a system for creating a live question and answer service with real experts facilitated by MILA–T. The inquiry orientation of MILA is different from settings like River City that attempt to create entire virtual environments in which students can simulate any inquiry task. Rather, MILA's approach to inquiry is to facilitate and direct execution of real inquiry tasks regarding the phenomenon within the software itself.

On the modeling side, MILA leverages a modeling framework called CMP (Component-Mechanism-Phenomenon) modeling, described in greater detail below. On the software side alone, however, the approach to delivering this modeling framework provides a natural set of constraints to sculpt students' reasoning even in the absence of a tutoring system. The structure of the interface pushes students to think about modeling in terms of interacting trends and changing variables by reframing any claims or explanations in terms of causal sequences of events characterized by changes to the state of the system. This structure is instructed early in the use of MILA, but the constraints of the software push students to think in terms of the types of models that MILA can create. Most importantly, MILA inserts the results of the inquiry activities directly into the modeling process by asking students to back up the claims of the model with evidence derived from their inquiry activities, tightening the link between modeling and inquiry.

As presently constructed, MILA has been successful in facilitating modeling of ecological phenomena in middle school classrooms. Although the primary targets of analysis here have been the effects of the tutoring system on top of MILA itself, prior analysis has identified the software's ability to drive inquiry and modeling even in the absence of such feedback. The semantics of the models in the software are presently such that it may be used in any inquiry and modeling task in the ecological domain, and ongoing research is examining additional ways to build a more robust toolkit for modeling ecological phenomena and providing more naturalistic feedback through simulations.

CMP Modeling

As referenced above, MILA is constructed specifically to facilitate CMP (Component-Mechanism-Phenomenon) modeling. MILA itself could use a different modeling framework, while CMP modeling could be used in the context of a different exploratory learning environment, separating out the contributions of MILA and CMP. MILA's precise instantiation and implementation of CMP principles provide some contributions outlined above, but the broader formalization of CMP modeling as an approach to teaching modeling and inquiry in schools stands as an independent contribution as well. CMP modeling is constructed off of Structure-Behavior-Function (SBF) modeling, a language for constructing models originally created to facilitate design in artificial intelligence systems (Prabhakar & Goel 1998; Goel, Rugaber & Vattam 2009), but more recently applied extensively to biologically-inspired human design (Vattam et al. 2011B; Wiltgen et al. 2011). Research has been performed applying SBF modeling in education both at the college level (Helms, Vattam, Goel, & Yen 2011; Yen, Weissburg, Helms, & Goel 2011) and in middle schools (Goel et al. 2013). This latter research in middle schools found that usage of SBF principles in learning about complex ecological systems increased students' deep knowledge about those systems (Goel et al. 2013) as well as facilitated deeper reasoning skills in science teachers (Sinha 2010). A major difficulty in SBF modeling was the complex modeling semantics and foreign vocabulary used in these early experiments. In response to these observed problems, CMP modeling was developed to provide a more understandable set of semantics for middle school students to describe ecological phenomena as well as a simpler modeling interface to construct these explanations (Joyner et al. 2012).

CMP modeling is not tied specifically to the semantics seen in MILA; while MILA uses nodes to represent trends that arise out of underlying components and properties, one could develop a system for CMP modeling that instead models the individual components as nodes with multiple properties directly invoked within the component. The research on a different extension to MILA, MILA–S, takes exactly this approach and is documented more fully below. Moreover, CMP modeling does not strictly have to take place in a software environment; it can be performed on paper as well. Earlier experiments with SBF modeling began by having students design models of ecological systems on paper using the SBF terminology and structure (Honwad et al. 2010). Similarly, CMP modeling may be leveraged outside an exploratory software learning environment in an exercise wherein students clearly articulate a target phenomenon to describe, describe the multiple physical components and their respective properties, and arrive at a mechanism explaining how a sequence of changes in properties led to the noted phenomenon. Thus, CMP modeling provides a robust and promising approach to teaching modeling separate and apart from its specific implementation within MILA.

Model of Inquiry-Driven Modeling

The existing literature from the learning sciences and technology community is replete with arguments supporting the desirability of an inquiry-based approach to science education (e.g. Schiff 1970; Keys & Bryan 2001; Gibson & Chase 2002) as well as the proper conductance of such open-ended, model-based, project-oriented, constructionist approaches to science (e.g. Clement 2000; Sins, Savelsbergh, & van Joolingen 2005; Schwarz et al. 2009). Schwarz et al. 2009 in particular provides a progression through the phases of learning about scientific modeling that students ought to complete, emphasizing educating students on the cycle of model construction, use, evaluation, and revision. Work has also been performed on the union between modeling and inquiry and how scientific modeling can play a critical role in driving the inquiry process (e.g. Jackson 1995; Stratford & Finkel 1996; Löhner, van Joolingen, Savelsbergh, & van Hout-Wolters 2005; Svoboda & Passmore 2013). Most notably, the Model-It project provides students with a tool for constructing scientific models in the service of authentic scientific inquiry (Soloway et al. 1997).

However, despite the considerable time put into researching the use of modeling and inquiry in educational settings, relatively little has been done on formalizing the process into a cohesive model of the target skill. There is an understanding and acceptance of the notion of model construction, use, evaluation, and revision at a high level, but not of the way in which scientific inquiry can drive, and subsequently be guided by, scientific modeling. In order to instruct students in this inquiry-driven modeling approach, this work began by formalizing the existing literature on this skill into a process flow for the investigation of a phenomenon and construction of a model of the phenomenon, including the possibility that the primary takeaway of the process is the rejection of the model as insufficient or inaccurate. This process was constructed according to the existing literature and supported by our own prior experience with developing exploratory learning environments to facilitate construction of scientific models in inquiry-based learning activities (Joyner et al. 2012; Joyner, Majerich, & Goel 2013; Goel et al. 2013). In addition to the articulation of this process, we also augmented the model of the process with a collection of errors we have commonly witnessed students committing in their inquiry-driven modeling behaviors in the past.

While this model was constructed based in large part on the literature on inquiry and modeling both in education and in professional practice, it is not claimed to be a validated model of the way in which inquiry-driven modeling is (or should be) performed by scientists in the real world. Rather, this model is claimed to be a desirable target for which to aim in interventions addressing students' modeling and inquiry skills. This articulation of a productive process of inquiry-driven modeling and the accompanying errors that students may commit in the course of the process should provide a useful starting point for others analyzing students' progress in learning this authentic process of scientific inquiry and modeling.

Guidance for Inquiry-Based and Discovery-Based Teaching

Although providing this model of a desirable inquiry-driven modeling process is a contribution to corners of the learning sciences & technology community that support inquiry-based, discovery-based constructivist teaching, there exist many critics of this approach. In many classrooms and in prior experiments along the same line of reasoning as this research, students complete inquiry-based, discovery-oriented projects with little or no guidance based on the belief that the greatest learning gains are seen in natural investigative experiences. Some theorists, however, argue that "minimal guidance during instruction does not work" (Kirschner, Sweller, & Clark 2006). They argue, among other claims, that students only begin to benefit from receding guidance when students achieve a rate of internal knowledge to provide guidance for themselves. Teaching this kind of knowledge and self-guidance, however, is very difficult in open-ended discovery-oriented endeavors because a significant level of guidance is required in the interim which is difficult for a single instructor to provide to multiple teams simultaneously; this echoes the well-documented "two-sigma problem" in education (Bloom 1984). To put these claims simply, students need to achieve a level of expertise to learn from discovery-based, inquiry-based activities, and that expertise is difficult to learn within the activities themselves.

We would argue that in many cases, inquiry and discovery are not merely the means to achieve certain learning goals, but rather are the learning goals themselves. Nonetheless, the idea that learning gains are not seen simply from the opportunity and subsequent completion of an inquiry-based learning activity is not controversial. In fact, MILA–T is designed with exactly this problem in mind; given an exploratory learning environment in which to participate in an inquiry activity, there still exists the opportunity for instruction and feedback to support the learning process. A tutoring system like MILA–T then can play the role of guiding instruction while preserving the elements of the activity that deliver either explicit gains to students' inquiry-driven discovery or gains to the underlying learning goals that inquiry-driven approaches attempt to accomplish.

This claim that a tutoring system like MILA–T can provide individual guidance to multiple students or teams simultaneously in a way a single teacher cannot is expected; it is fundamental to the goal of tutoring systems, to bring immediate and individualized feedback to many students concurrently. However, what is particularly notable about MILA–T in light of this criticism is that MILA–T addresses a domain in which students are commonly participating in inquiry-based learning already. Inquiry-based approaches are rarely used in mathematics or many other domains in which tutoring systems have been successful in the past. MILA–T not only operates in an inquiry-based learning environment, but in fact the inquiry itself is the targeted learning goal of the tutoring system. Thus, MILA–T plays precisely the role of guiding instruction as advocated by Kirschner, Sweller, & Clark while preserving the inquiry-based nature of the learning activity.

Discussion

The final discussion of this work focuses on three different areas for continued application of the tools developed and lessons learned throughout this study. First, this discussion explores two possible directions for future research within this project: examining the hypothesis regarding the interplay between a tutoring system and a teacher in a classroom and developing alternate ways to provide students feedback on their progress in developing explanations for phenomena. Second, this discussion describes the way in which the lessons and design principles learned throughout this study have been applied more recently in a different domain. Third, this discussion mentions three new collaborations spawned by this work exploring how to apply these tools and principles to new audiences at larger scales.

Future Directions for Research

One of the significant takeaways of the previously-described research is the hypothesis that the mechanism by which the presence of a tutoring system led to learning gains in the classroom may be through its impact on the way in which the teacher conducted the classroom generally, freeing up additional cognitive resources to provide deeper feedback. This hypothesis provides a valuable next step in evaluating the cause of the gains seen in this study. Secondly, while in this study feedback was provided to students as part of interaction with a tutoring system, there exist alternate ways in which students can receive feedback and improve accordingly. One such way has already been explored: develop a system within MILA whereby students can directly invoke simulations that depict the way in which the system would operate if their model of the system was accurate. The differences between this simulated system and the actual system then become feedback for students to improve their model more going forward.

Intelligent Tutors in a Classroom Setting

Throughout this analysis, a common thread has been the proposition that the results seen here could be explained by the effect that the presence of a tutoring system has on the classroom as a whole rather than the individual students or teams using it. The prior analysis showed that there existed trends in the Experimental group that are difficult to tie directly to the feedback provided by the tutoring system. While it is true that the tutoring system aims to provide students with a greater understanding of the nature and role of evidence in supporting explanations, the tutor lacks the ability to correct students' flawed evidential justifications. Therefore, either students in the Experimental group learned a better understanding of the nature of evidence without actually receiving targeted correction on the evidence they have already written, or something else was responsible for the presence of these learning gains solely in the Experimental group. In the above sections, it was proposed that the primary role of the tutoring system was to receive and address lower-level requests for assistance, allowing the teacher in the classroom to field fewer questions and devote more cognitive resources to observing the classroom as a whole. This also would suggest that when the teacher did give students' feedback, it was after they had already exhausted the feedback available from the tutor, meaning that the specific feedback was of a higher caliber (e.g. "This piece of evidence is not relevant to this claim" or "This piece of evidence is insufficient."). This proposition was further supported by the interviews with the educators in the classroom; two educators reported that they were aware of answering fewer simple questions in classes in the Experimental group, and the third reported that she had difficulty at first allowing the tutors to play her typical roles for her. Thus, there are sufficient data to justify exploring the nature of teacher feedback in the classroom during engagement with a metacognitive tutoring system.

The data gathered as part of this analysis were not sufficient to draw strong conclusions about these interactions, however. In order to more rigorously test the hypothesis that the presence of a tutoring system allows teachers to focus on giving higher-level, more advanced feedback, two studies may be conducted. First, more simply, if the benefit of the tutoring system in this exercise was largely based on the feedback that teachers gave, then the teachers may be removed from the classroom entirely for a subsequent iteration, and interaction between teams would be limited to remove the presence of a "classroom-level" dynamic. The Control and Experimental groups would remain largely the same, with each group presented the same materials as well as standards for successful completion of the exercise. If the gains seen in this study were replicated in this follow-up study, then the gains can be attributed solely to the system consisting of the individual team and the tutoring system. If, on the other hand, the gains disappear in this follow-up study, then the benefit of the tutoring system was connected to the way in which it changed the classroom as a whole.

However, that method of conducting a follow-up study would involve intentionally removing an element that is presently believed to be beneficial. An alternate manner of conducting the study would instead involve retaining and examining exactly the feature presently believed to potentially be at the core of the learning gains seen in the Experimental group in this study. This would involve running effectively the same experiment in design, but carefully tracking the actions of the teacher during the intervention. The teacher would wear a microphone and have all interactions recorded for subsequent analysis. Simultaneously, an observer would sit in the classroom and monitor the proceedings using an observational protocol like the Reformed Teaching Observation Protocol (RTOP; MacIsaac & Falconer 2002) to track differences between the Conditions at the classroom level. After gathering these data, the questions asked of the teacher and the teacher's responses would be coded to check for differences in the types of questions asked of the teacher and the types of responses given. Based on the theory suggested by this prior analysis, in that study we would hypothesize that in classes with a tutoring system, the teacher would deliver a greater number of deep pieces of feedback and spend less time providing basic or superficial feedback. We would also hypothesize that there would be a greater frequency of feedback given to the class as a whole based on the teacher's observation of ongoing trends.

While this analysis has a special relevance to metacognitive tutoring and metacognitive tasks given the aforementioned difficulty in observing and reacting to metacognitive development, it also extends to general intelligent tutoring as well. In more constrained domains, such as Algebra or Physics, there is often an expected set of common mistakes that an intelligent tutor can anticipate and correct. Such an idea is so common that the community has moved toward extending to teachers the ability to set up such tutors because the limiting factor is the limited amount of content (Aleven, McLaren, Sewall, & Koedinger 2009). However, even in these settings there may often remain patterns of errors or misconceptions that have never previously been observed, or students who have difficulty overcoming their errors even with the aid of such tutoring systems. In these settings as well, the presence of a human teacher or tutor may help overcome these final difficulties while also leveraging the ability of the intelligent tutoring system to take care of the vast majority of expected misconceptions and errors.

Ablation of Functional Roles and Features

The takeaway of the previous analysis is that the presence of a metacognitive tutoring system during engagement with an activity in scientific inquiry leads to positive results in students' attitudes toward science and explanations of the target phenomenon. However, based on this study, this conclusion only makes a claim about the presence of a metacognitive tutoring system compared to the absence of one. The possibility remains that the gains seen here should actually be more narrowly attributed to certain functional roles played by the tutors in MILA–T rather than the tutoring system as a whole. Alternatively, the possibility also remains that the gains present here could be more widely attributed to any tutoring system present in the system under the same experimental design, whether oriented toward metacognitive ability, cognitive ability, or content knowledge. Similarly, it may also be argued that the gains present here are due solely to superficial factors of the tutoring system, such as the simple presence of feedback (regardless of the nature of the feedback) and the humanization of the tutors through human avatars.

To more specifically note that specific origin of the positive effects seen in this study, a more robust set of tutors ought to be created. In addition to the five metacognitive tutors present in MILA–T, additional tutors playing the same five functional roles should be constructed to target cognitive ability and content knowledge. Within cognitive ability, the tasks to teach would likely involve controlling for particular variables in a simulation, analyzing numeric results from an experiment, or gathering information from existing literature (although arguments could be made that any of these may be classified as metacognitive skills). Within content knowledge, the tutors ought to be equipped with the ability to specifically evaluate the ecological accuracy of the claims made by students' models, as well as to provide students targeted information from the literature based on the current claims and structure of their explanations. Alternate versions of the system should also be constructed removing the faces and names of the tutors of MILA–T.

After constructing these additional tutors, an ablation study would be run to vary the feedback that students received. Some students would receive only feedback of certain kinds, like metacognitive, cognitive, or content knowledge. Other students would receive multiple kinds of feedback, but only from certain functional roles, like the Guide or the Mentor. Still others would receive the full range of available feedback, but without the avatars or names, while a final set would be given a "placebo" set of tutors that provides no feedback but appears to monitor and react to students' activity anyway. Based on the analysis of these data, questions could be answered about the specific gains seen from specific types of feedback, as well as the specific gains seen from particular functional roles. Based on the results here, the likely hypothesis would be that targeted feedback on the current weaknesses and areas for improvement in students' current explanations from the functional role of the Critic would be most beneficial, and that such feedback would be beneficial in different ways across all three dimensions.

Alternate Methods of Feedback

The proposed study above concludes by hypothesizing that the most useful functional role in an ablation experiment would be the Critic. The ultimate reason for this is that the Critic plays a fundamentally important functional role with regard to any kind of investigation, discovery, or invention process: feedback. While all the tutors provide feedback of some kind, the Critic provides feedback that is most directly associated with identifying and repairing the weaknesses in a team's current explanation. This feedback process is a fundamental part of the modeling process documented earlier, featuring model construction, use, evaluation, and revision (Schwarz & White 2005). In this workflow, the Critic in many ways fulfills the "use" portion of the cycle; teams use their model examine the weaknesses in their current understanding, and the Critic provides feedback on what the results of that examination ought to be. In this way, the Critic helps improve teams' models by directing attention to the weaknesses, while also helping develop students' metacognitive understanding of modeling by articulating the value and purpose of such evaluation.

However, feedback from an intelligent tutoring system is only one of multiple possible ways to simulate this feedback process in a classroom exercise. Within experiments using MILA, an alternative approach has been attempted that has shown promising early results. This approach does not involve a metacognitive tutoring component, but it is relevant to the analysis here in that it presents an alternative method for providing feedback to students in service of the broader inquiry-driven modeling process, as well as part of the full picture of the intervention.

This alternative system is called MILA–S, for MILA–Simulation. In MILA–S, models that students develop are used to directly invoke a NetLogo simulation (Wilensky 1999). Whereas the NetLogo simulations used in the intervention involving MILA–T served as stand-ins for the real world and did not change, simulations generated by MILA–S are direct compilations of the claims made by students' models. The role of these simulations is to give students feedback on the way in which they system would behave if their current model of the system is accurate. If the simulation of their current understanding of the phenomenon does not match the real events in the phenomenon, then it stands to reason that a portion of that current understanding is insufficient or incorrect. The simulation, then provides teams with information on whether or not their current understanding matches the real system, as well as the specific places where the difference arises. This is a feedback process similar to the role played by the Critic in MILA–T; the difference here is that the feedback is naturalistic and raw rather than intelligent and selected. As a result, however, this feedback can be received without any prior work being put into developing the simulation; the Critic, on the other hand, needs some understanding of the activity in which students are engaging to supply feedback. A full description of the way in which MILA–S takes MILA models and compiles then into executable NetLogo simulations can be found in Joyner, Goel, & Papin 2014. Prior work on connecting conceptual models to NetLogo simulations can be found in Vattam, Goel, & Rugaber 2011.

During the intervention described in the preceding analysis, teams in the Gifted classes were given access to MILA–S as part of a final one-day exercise to further investigate the fish kill in Lake Clara Meer. This exercise took place after these teams had already submitted their final Learning and Transfer models and completed the post-test and survey, and so this activity did not influence any of the prior results. During this activity, teams were given 45 minutes to model the relationships in Lake Clara Meer in a way that would give rise to the observed phenomenon in the simulation. Within this short exercise, students were seen constructing initial models, generating simulations, and revising the models based on observed instances where the simulation did not match the expected results. Students were also seen using their models to reflect on the real systems; for example, one team constructed a model that accurately recreated the fish kill, but decided to revise the model because they believed it ignored important components from the real system. Other interesting trends were observed, such as teams engaging in a nested modeling process wherein the simulation became a model that was used, evaluated, and revised, and whose revision process led to feedback that informed the evaluation and revision of the conceptual model. A broader description of the patterns of interaction observed during this activity can be found in Goel & Joyner 2014.

In order to facilitate this expanded functionality, MILA–S models are augmented with an additional layer of information on top of what is typically present in a MILA model. An example of this is shown in Figure 29 below. Rather than representing trends, each node in this interface instead represents an individual component, such as Oxygen, Algae, Sunlight, or "Fishies". Components are annotated with an assertion that they are either biotic or abiotic, and that differentiation determines the variables associated with the component; biotic organisms possess populations, lifespans, birth rates, and energy levels, while abiotic substances possess only quantities. Additionally, relationships between biotic and abiotic components are further explained as instances of different relationship prototypes, like 'eats', 'produces', or 'destroys'. These relationship prototypes provide information to MILA–S that is used to compile the simulation.

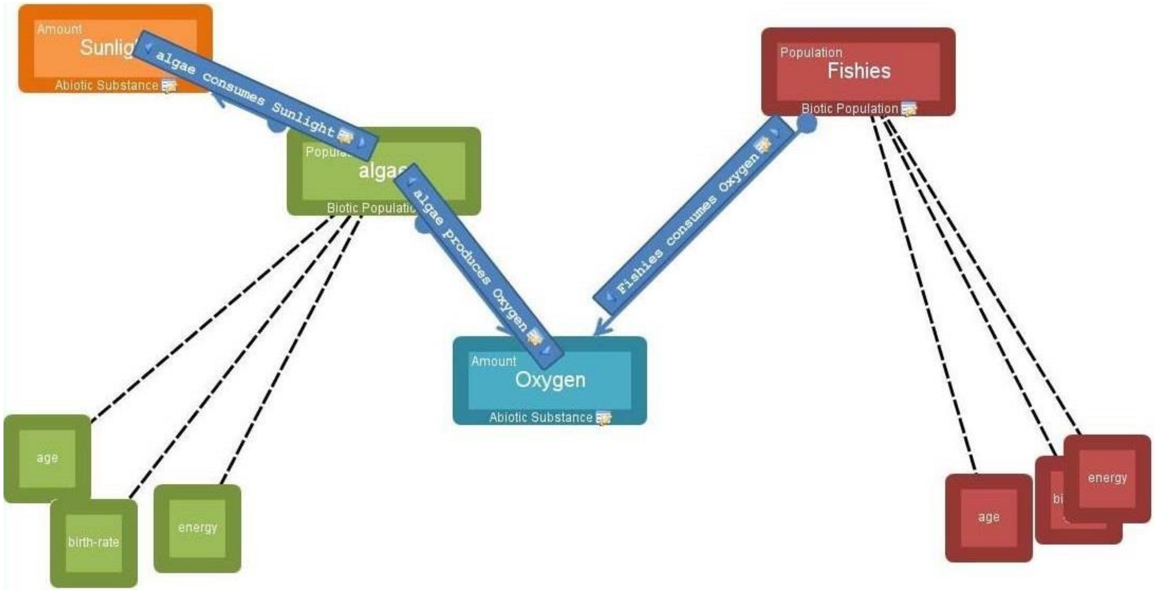


Figure 29: A MILA–S model produced by a team during the project at the end of the intervention, originally published in Joyner, Goel, & Papin 2014.

At present, there does not exist a "MILA–ST", a system unifying the generated simulations of MILA–S with the metacognitive tutoring capabilities of MILA–T. Early thought has been paid to how a set of metacognitive tutors could be developed within MILA–T that may provide information useful during engagement with MILA–S. Planned development of MILA–S will focus on improving the system's ability to accurately simulate ecological phenomena by building in additional ecological content knowledge and a more complete ontology of the types of components, variables, and relationships necessary to fully model and understand an ecological phenomenon.

Application to MOOCs

Most recently, the results and lessons learned throughout this project have been applied in a radically different domain. In February of 2014, we began the development of a MOOC – a massively open online course (Pappano 2012) – on Knowledge-Based AI, a course taught at Georgia Tech by Prof. Goel for the past several years. Initially, this course was developed to be offered as part of the Georgia Tech Online Masters of Science in Computer Science (Lewin 2013), with plans to offer it as a freely available

MOOC in the future. In the past, significant criticism has been leveled at MOOCs for their failure to recreate desirable pedagogical experiences, instead relying on the faulty lecture-oriented model to reward only the most self-driven learners (e.g. Knox et al. 2012; Martin 2013; Ross et al. 2014; Guzdial & Adams 2014). In developing an online course for Knowledge-Based AI, our goal from the beginning was to leverage the lessons learned within this project and the communities from which this project sprung. Although much could be written about the variety of learning strategies we have attempted to leverage in the development of this course, most relevant to the research presented in this document are the explicit focus on metacognitive development and the inclusion of intelligent tutoring-style individualized feedback wherever possible.

Metacognitive Development in a MOOC

As a course, Knowledge-Based AI relies heavily on the development of metacognitive ability in students. Much of the course emphasizes a feedback cycle not unlike the process of modeling seen in the previously-described intervention. In Knowledge-Based AI, we teach students to use our understanding of human cognition to develop artificially intelligent agents that perform the way in which humans do, and then to subsequently use the performance of those agents to reflect on the nature of human cognition. In this way, the agents can be thought of in many ways as models of our understanding of human cognition, and by using the agents in simulations or to perform tasks, we may gather information that allows us to evaluate and revise our model of human cognition. In this way, a major learning goal of Knowledge-Based AI is the ability to think about our current understanding of the human mind, to think about the inner workings of an artificially intelligent agent, and to use each to reflect on the other. In this way, Knowledge-Based AI is a richly metacognitive course.

Because of this learning goal, the course is designed from the ground up to emphasize the instruction of metacognition to the students in the course. After the customary introduction and lesson outline to set expectations and connect material to past topics, each lesson in the course starts with an example for students to complete. These are often duplications of basic tasks that individuals complete in the real world every day, such as understanding a simple sentence or finding the best container to use to transport soup. This example sets up the initial reflection of the lesson, examining the process by which we humans complete complex tasks with relative ease. This example also typically serves as the task which we design an agent to accomplish, drawing a close connection between the agent and the human mind based on their mutual ability to solve the task presented at the start of the lesson. As the lesson progresses, students step through the reasoning behind solving the problem as the agent would do so, thus consistently reflecting on how their mind is able to accomplish the necessary steps toward problem solving that the agent must be equipped to do. At the conclusion of each lesson, students are asked to reflect on what they themselves learned, what concepts they found difficult, and what connections they could draw to earlier course material. This process attempts to provide an explicit opportunity for self-regulation through the course, although the data will have to be analyzed to determine whether or not students actually take the opportunity to do so given the inability of the system to monitor for earnest responses.

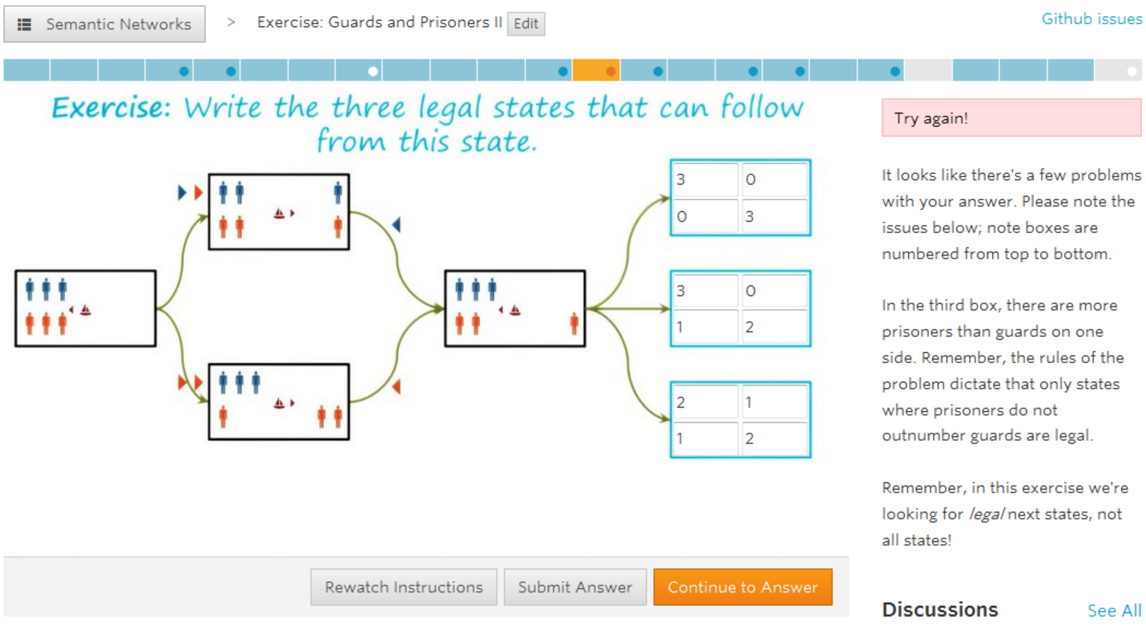
This focus on metacognition is not limited to the nature of the lesson planning. In many ways, the lesson plans are the least interesting part of this endeavor because nothing they provide is MOOC-specific; the exact same lesson plans could be developed and applied to an in-person class. However, certain production elements throughout the course are constructed to facilitate metacognitive development in a way that would be difficult in a traditional classroom. Throughout the course, there are two individuals interacting, the course developer and the professor. In many instances, the two interact in a way that is intended to demonstrate to students exactly the type of metacognitive abilities they ought to develop on their own. The professor, for example, will often ask the course developer to answer the exercises that students are asked to complete as well.

The course developer, in turn, will provide an answer, but will also move through an entire explanation of the exact thought process that went into the answer that the student should learn to mimic as well. This often involves specifically anticipating certain errors or misconceptions that the student might make and addressing how the course developer avoided those errors. In this way, the course developer demonstrates to students exactly the type of thought processes they ought to learn specifically in the context of the exercise that they, too, just completed. While it would be possible to duplicate this dynamic in some ways in an in-person classroom, the nature of the scripting process, the possibility for interaction between two "characters", and the ability to directly build on an exercise that students have just completed all make this approach uniquely well-suited to the pre-recorded medium.

Individualized Feedback in a MOOC

Intelligent tutoring is, ultimately, a matter of feedback. Based on the activity and input of the student, the tutor provides some feedback to help the student improve and more closely approximate the target skill. In most MOOCs, the opportunity for practice with real feedback is minimal or non-existent; in fact, it is exactly this lack of interaction and learning by doing that is fundamental to many of the criticisms of MOOCs. Without this interaction, learning is entirely instruction-based with the learner playing little role in actively participating in the process unless they themselves choose to break out of the system and apply their skills elsewhere. This lack of interaction is partially responsible for the criticism that MOOCs are only successful for the most self-driven learners; there is nothing to hold the attention or engage learners that are not already naturally engaged with the material.

The Udacity development platform provides a medium for designing interactive exercises that can be graded by a script. Exercises can consist of any number of textboxes, radio buttons, and checkboxes, all of which can be leveraged dynamically in the code. At a purely functional level, this provides the infrastructure necessary for designing interactive exercises. An example of such an exercise from the Knowledge-Based AI class is shown below.



Exercise: Write the three legal states that can follow from this state.

3	0
0	3

3	0
1	2

2	1
1	2

Try again!

It looks like there's a few problems with your answer. Please note the issues below; note boxes are numbered from top to bottom.

In the third box, there are more prisoners than guards on one side. Remember, the rules of the problem dictate that only states where prisoners do not outnumber guards are legal.

Remember, in this exercise we're looking for *legal* next states, not all states!

Figure 30: An exercise from the Knowledge-Based AI course. In this exercise, students are learning to use a semantic network to create a formal representation of a problem state. As part of this, they have previously seen the creation of some nodes of the semantic network, and are now asked to create the possible next states. In doing so, they participate in the reasoning process that helps make clear how the semantic network representation is being used to support the problem solving process.

Given this infrastructure, a small set of simple quizzes are most common. Oftentimes, multiple choice quizzes are written with one or more correct answers. Occasionally, students may be given feedback on why their current selection on these quizzes is wrong, but oftentimes students are instead asked simply to try again. Similarly, in free-response quizzes, students most often have their answer checked by a simple string comparison. If the string given by the student does not match the "correct" answer to the exercise, the student's response is marked wrong and a token hint is provided. While there do exist quizzes that go beyond this model, this model reflects the majority of quizzes seen in other classes.

Based on our prior experience with intelligent tutoring systems, however, our approach to these exercises was from the start motivated by the idea of developing "nanotutors" for each type of exercise. These tutors are motivated by the desire to provide useful feedback at every stage of the problem solving process, no matter how far the student is from the correct answer. In this way, feedback is mapped according to a series of bridges from a certain level of ability and correctness to the next. This motivating set design is depicted visually in Figure 31.

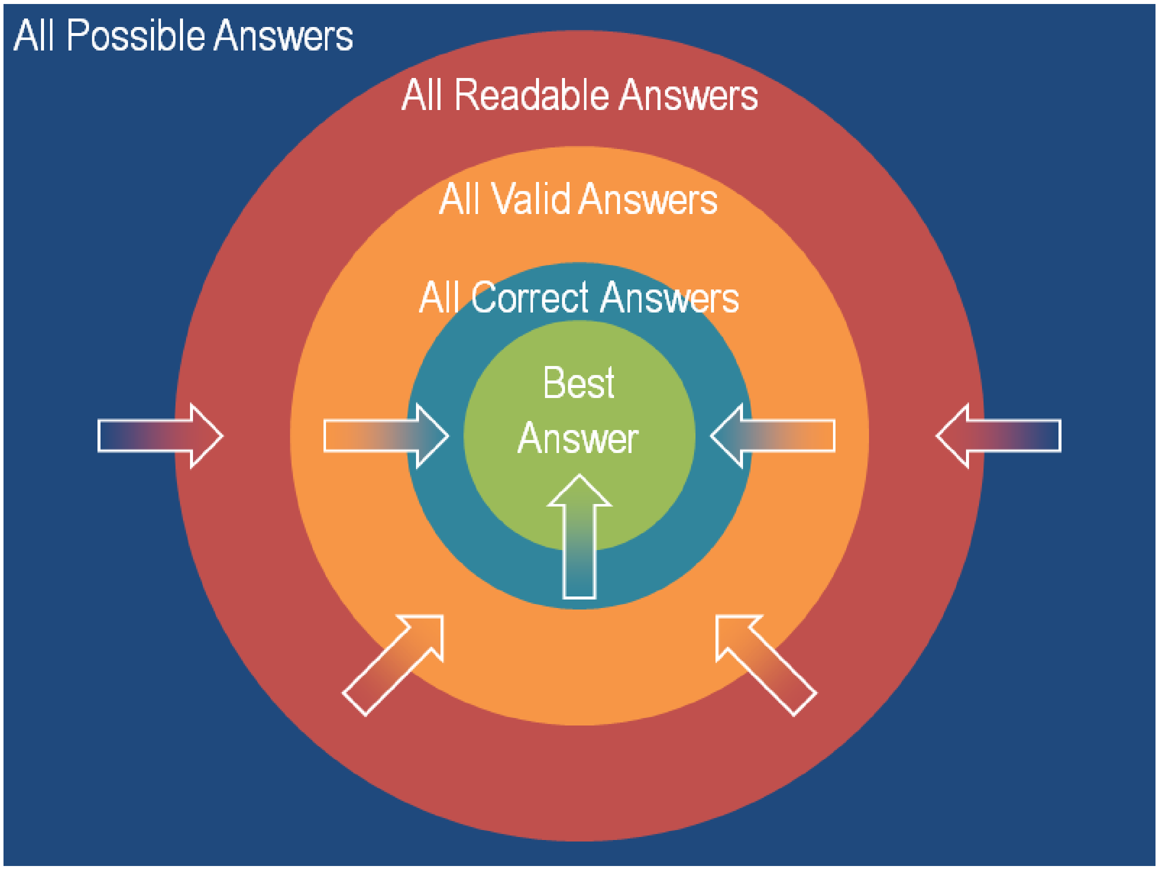


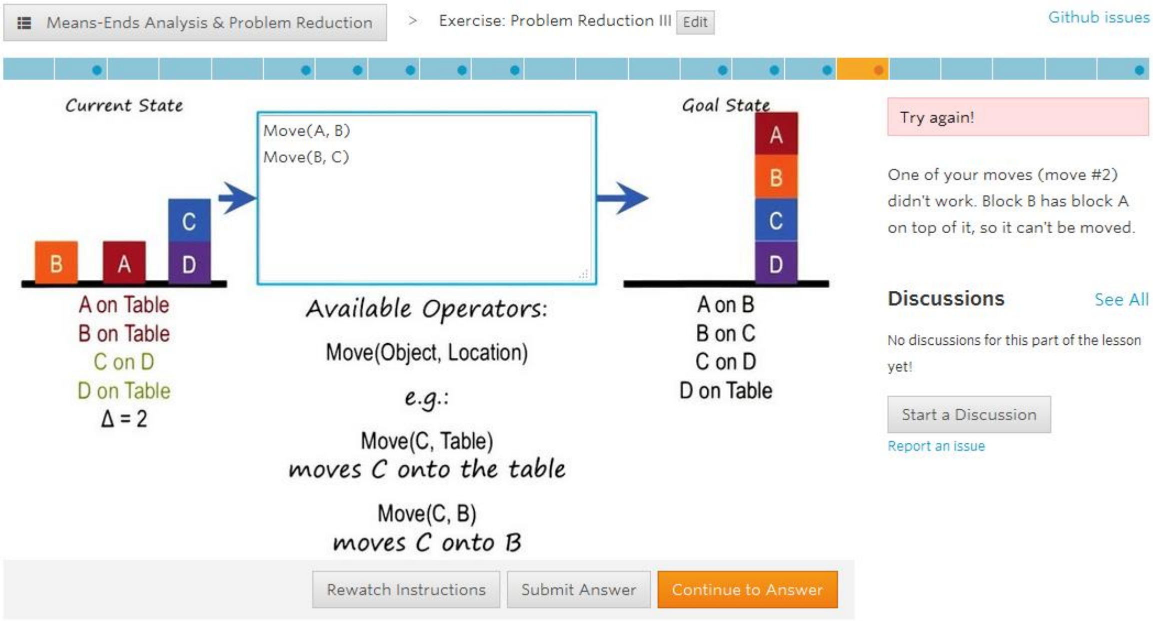
Figure 31: A visualization of the motivating design principle behind exercises in the Knowledge-Based AI course.

The nanotutors in these exercises play the same role as the Critic in MILA–T: they look at the student's current understanding of and answer to the problem and provide guidance on how to reach the next level of understanding. This starts first by taking a look at the student's answer and ascertaining whether it is a readable answer for this problem. For example, in the exercise shown in the Figure 30 above, students are asked to fill out four numbers that together represent a state of the problem. If students were to put in letters, their answer would fall into the broad blue rectangle outside the red circle.

The nanotutor would inform the student that the only valid answers to the problem are numbers, moving them from the blue space to the red circle.

Once the student has put in an answer that the nanotutor can understand, the nanotutor can begin to map out its understanding of the student's answer and test for its correctness. In some (but not all) exercises, this first involves a test of validity. While the answer may be readable, it may not be valid. For example, in the exercise shown above, one of the states does not obey the rule of the problem that the bottom number can never be larger than the top number (unless the top number is 0). For this reason, this is not a valid state. As shown in the bottom right, the student is receiving feedback on this fact; their answer falls into the red circle of readable but invalid answers. They are, thus, given instruction on how to proceed to the next stage by ensuring that only valid states are included. As the problem continues, the student may arrive at a valid answer that nonetheless is not correct for the problem, such as including the same state twice in the exercise above; doing so would be valid, but not correct. Finally, in many exercises, there is a set of answers that are technically correct but are not optimal. For these answers, students are alerted that their answer is correct, but also that there answer still differs from the tutor's in some ways. For example, in a subsequent exercise, students finish solving the problem given in the above exercise and input the number of moves required. If the number of moves that students enter is possible (an odd number greater than or equal to 11) but sub-optimal (greater than 11), the nanotutor tells students that while their answer is correct, a better answer does exist, along with a tip on how to arrive at the better answer.

To illustrate this further, another example is given in Figure 32 below. Here, students are asked to complete a problem of rearranging blocks, a common problem in AI. This problem is used to reflect on how an AI agent can detect and avoid conflicts in advance, prioritize and select actions, and evaluate different paths from its current state to its goal state. Students here write out the operators that the agent might execute to arrive at the goal state. Here, in the first phase, the nanotutor checks to see if the input is valid; each line of the box should be a move command with two parameters, and if it is not, the student is informed of the invalid line or lines. After confirming that all of the commands are readable, the tutor begins executing them. If it arrives at a command that it cannot execute within the rules of the problem, as seen above in the attempt to move B while A is on top of B, the answer is ruled readable but invalid, and the student is given the opportunity to revise the answer to add validity. If all the moves made within the problem are valid, then the nanotutor checks to see if the answer is actually correct; if the blocks are in the desired arrangement at the end of the process, it tells the student that their answer is correct, and if they are not, it tells the student the exact arrangement of the blocks at the conclusion of the moves.



```
Move(A, B)
Move(B, C)
```

A on Table
B on Table
C on D
D on Table
Δ = 2

Available Operators:
Move(Object, Location)
e.g.:
Move(C, Table)
moves C onto the table
Move(C, B)
moves C onto B

A on B
B on C
C on D
D on Table

Try again!

One of your moves (move #2) didn't work. Block B has block A on top of it, so it can't be moved.

No discussions for this part of the lesson yet!

Figure 32: Another exercise from the Knowledge-Based AI class. In this exercise, students complete a planning exercise where they examine how an agent would be able to reconcile potentially incompatible goals.

There are two crucial elements to the design of both these exercises. First, as described by Figure 31, at every given step of the problem-solving process students are given feedback on how to get closer to the answer. This mimics the role of the Critic in MILA–T, which gave students at any stage of the project feedback on how their model could be better. The student is never left guessing as to the best way to proceed further in the problem, and in most cases, the student is also specifically informed of the reason or mechanism behind the current error as it relates to the design of the agent. The second crucial element to note here is that students are actively engaging in the exercises that are themselves the target concepts of the lessons. In the first exercise, students are learning how an agent might use a semantic network to generate potential next states and test them for usefulness, and to do so, they actually generate the states themselves. In the second exercise, students are reflecting on how an agent would select and prioritize actions, and to do so, they select and prioritize actions according to the same heuristics that the agent uses. All along, they are given explicit, targeted, individualized feedback specifically on the weaknesses of their current answer and the way in which the answer can be improved.

Evaluation and Summary

The online course described here officially launched to a closed audience of 200 students of the Georgia Tech Online Masters in Computer Science program in August of 2014. Analysis of the data generated by students in the class is underway. The ultimate goal of this analysis is to determine the ways in which students interacted with this feedback, such as iterating over the answer, submitting one answer and giving up, or skipping the exercise entirely. Present analysis, however, has covered students' perceptions of the tool and its usefulness in learning the material. When asked for general feedback on the class as a whole, one student replied:

> *"On the Udacity quizzes I can really tell that you put in an effort to give back meaningful responses (way to be awesome at AI). They're seriously the best responses I've ever seen in an online class (I've taken about 10)."*

A second student responded similarly:

> *"I have to say that the videos combined with the puzzle exercises have made the concepts much, much clearer."*

A third student echoed those comments as well:

> *"There are tons of exercises throughout the lecture and good descriptions about those exercises about how they got their answers. I like this a lot."*

Continued analysis of the data generated by students in the class will provide a more comprehensive look at the way in which the nanotutors supplied throughout the course lessons were leveraged by students and the way in which those patterns of interaction correlated with perception of and performance in the course as a whole.

Ongoing Collaborations

The scope of this research has focused narrowly on teaching inquiry and modeling abilities to middle school students. However, the broader goal of this research program is to address the full scientist life cycle: secondary education, higher education, amateur scientists, and professional scientists. One of the key motivations behind this research was to bring an authentic scientific experience to secondary school students, but that authentic scientific experience needs tools to help facilitate the same kind of inquiry and modeling that we have taught to students in the first place. An ongoing goal of this research program is thus to support scientists at every level, training students at the secondary and college levels and aiding scientists at the citizen and professional levels. Toward this end, this dissertation will close by describing three ongoing endeavors to expand this research to these audiences: a collaboration with Georgia State University's College of Education to further help secondary school students; a collaboration with Georgia Tech's School of Biology to help college-level students; and a collaboration with the Smithsonian Institution's Encyclopedia of Life to support citizen and professional scientists.

Georgia State University's College of Education

In recent years, researchers at Georgia State University have examined many of the same issues present in the work presented here. Specifically, they have examined the challenges present in inquiry-based education (Renken, Carrion, & Litkowski 2014), the different kinds of evidence used in constructing science understanding (Goldin, Renken, Galyardt, & Litkowski 2014), and the role of computer simulations in science learning (Renken & Nunez 2013). Recently, this has taken the form of a project called Choose Your Own Science Adventure (CYOSA), a structured activity wherein students virtually investigate a natural phenomenon by gathering data, constructing an understanding of the mechanism underlying the phenomenon, and providing evidence in support of their mechanism (Peffer 2014). CYOSA differs from MILA in that it is significantly more structured; students navigate a predetermined set of dialogs and bits of information rather than explore a system openly. In this way, CYOSA offers a powerful complement to MILA. Feedback from the teachers in the study reflected the need for more novice-level help, and while the tutors aim to provide that guided instruction, the open interface of MILA makes it difficult to perfectly understand students' abilities. CYOSA provides an opportunity to create novice-level guidance in a more structured environment before transitioning students to MILA's more open-ended inquiry environment.

The proposed collaboration with Georgia State University aims to use CYOSA to provide novice-level instruction and scaffolding to complement the open-ended inquiry and modeling facilitated by MILA. Middle school life science students will begin by using CYOSA to receive highly structured information on the course of modeling and inquiry; during this time, a metacognitive tutoring system like MILA–T will still operate to emphasize, instruct, and demonstrate metacognitive processes during the structured inquiry task. Once students have demonstrated an understanding of the structured task, they will be transitioned to the unstructured task with MILA, while the metacognitive tutoring system helps bridge the gap between the two environments. At the conclusion of this process, students will tackle an even more unstructured problem without this tutoring system (and potentially outside the software altogether) in order to demonstrate the extent of their learning during the unit. Proposals for this research are currently under development.

Georgia Tech's School of Biology

The research described in this dissertation has focused on teaching metacognition, inquiry, and modeling to middle school students; however, these are important abilities for college students as well. Can the research developed here be transferred to college-level ecology classes? A collaboration with the Georgia Tech School of Biology aims to address this question, specifically building on the MILA add-on MILA–S described above on page 242. The learning goal of this collaboration is to teach students to use multiple kinds of modeling to improve their explanations of a system. More specifically, this collaboration aims to equip students with the ability to easily use conceptual models and simulation models together, improving their inquiry process and their explanations of the target phenomenon. Unlike the research on MILA and MILA–T, MILA–S is not meant solely as a teaching tool either; its ability to compile conceptual models into simulation models provides a practical function separate from demonstrating the process. While the learning goal is to teach students conceptual and simulation modeling as part of the inquiry process, the tool continues to play a role even after students master the ability to make investigating and simulating a phenomenon easier.

In using this tool, students will address the same phenomenon as the experiment in middle school classes: the fish kill in Lake Clara Meer. The goal of the activity will be to construct a model that (a) explains how the fish kill occurred, (b) is consistent with the circumstances surrounding the fish kill, and (c) is consistent with current ecological theories. The simulation compilation provided by MILA–S will help students evaluate whether the first goal has been accomplished, while traditional inquiry activities will provide information on whether the second and third constraints have been satisfied. The initial experiment with college students will be exploratory, observing students' inquiry and modeling behaviors and evaluating the effectiveness of the tool, the activity, and the students. Subsequent studies will examine the specific roles played by modeling, simulation, and tutoring. This exploratory study will begin in Spring 2015.

Smithsonian Institution's Encyclopedia of Life

As described above, while the goal of MILA and MILA–T are primarily to teaching inquiry and modeling, MILA–S takes this goal one step further and aims to aid real-world inquiry and modeling as well by providing an accessible way investigators to translate their present explanations into executable simulations. More information on this system can be found under the Alternate Methods of Feedback section on page 242. In ecological research, "citizen scientists" (that is, unpaid volunteers performing inquiry as a hobby) play a major role in documenting local phenomena and supporting real scientific practice (Oscarson & Calhoun 2007; Cohn 2008; Silvertown 2009; Howard & Davis 2009). In recent years, efforts have been made to equip these citizen scientists with tools to improve their ability to do real, rigorous scientific research (Stapleton, Smith & Hughes 2005; Reddy et al. 2007; Luther et al. 2009; Graham, Henderson & Schloss 2011). One of the challenges for citizen scientists, however, is typically the lack of programming expertise; scientists often use simulations of natural phenomena to test understanding and make predictions, but citizen scientists typically lack the ability to develop these simulations. MILA–S allows citizen scientists to have access to this simulation capability by directly compiling the more accessible conceptual models into executable simulation models (Joyner, Goel & Papin 2014).

The Encyclopdia of Life is a massive database of information about world ecology and biodiversity (Walter 2014). It contains millions of web pages of literature about the interactions of species, environments, and ecosystems. One of its main goals is to open up scientific knowledge to the masses, including citizen scientists who can significantly benefit from access to enormous amounts of scientific knowledge to use to examine their own local ecosystems. The collaboration with the Smithsonian Institution will combine the simulative capabilities of MILA–S with the content knowledge of Encyclopedia of Life. Using the two together, scientists will be able to construct executable models of their local ecosystems, leveraging the relationships and behavior patterns articulated within the Encyclopedia of Life. Work has also been considered to use natural language processing in conjunction with Encyclopedia of Life to automatically generate conceptual models of ecological phenomena that could then be passed to MILA–S to generate executable simulation models. Combined with AI like IBM's Watson, this research direction could ultimately lead to an agent that can autonomously read in simple data from the real world and relationships from Encyclopedia of Life and generate comprehensive models and simulations of ecosystems without any human intervention whatsoever.

APPENDIX A: ATTITUDINAL SURVEY

The attitudinal survey featured here is a combination of selected constructs from the TOSRA (Test of Science Related Attitudes, Fraser 1981) and mATSI (Modified Attitudes Towards Science Inventory, Weinburgh & Steele 2000). Combined, it measures five constructs: interest in careers in science, desire to do science, perception of the science teacher, perception of the science teacher, and anxiety toward science.

	Strongly disagree	Disagree	Unsure	Agree	Strongly agree
Grades are very important to me.	□	□	□	□	□
A job as a scientist would be interesting.	□	□	□	□	□
I have a good feeling toward science.	□	□	□	□	□
Working in a science laboratory would be an interesting way to make a living.	□	□	□	□	□
I would like to be a scientist when after I complete my school.	□	□	□	□	□
Science teachers are willing to give us individual help.	□	□	□	□	□
Science is something that I enjoy very much.	□	□	□	□	□
Sometimes I read ahead in our science book.	□	□	□	□	□
I feel tense when someone talks to me about science.	□	□	□	□	□
When I complete my school, I would like to work with people who make discoveries in science.	□	□	□	□	□
A career in science would be dull and boring.	□	□	□	□	□
I would dislike being a scientist after I complete my school.	□	□	□	□	□
Science teachers present material in a clear way.	□	□	□	□	□
It makes me nervous to even think about doing science.	□	□	□	□	□
I would dislike a job in a science laboratory after I leave school.	□	□	□	□	□
A job as a scientist would be boring.	□	□	□	□	□
I would dislike becoming a scientist because it needs too much education.	□	□	□	□	□
I like the challenge of science assignments.	□	□	□	□	□

When I hear the word *science*, I have a feeling of dislike.	□	□	□	□	□
I would like to teach science when I leave school.	□	□	□	□	□
It scares me to have to take a science class.	□	□	□	□	□
Science is one of my favorite subjects.	□	□	□	□	□
Science teachers make science interesting.	□	□	□	□	□
I have a real desire to learn science.	□	□	□	□	□
I would like to do some extra or un-assigned reading in science.	□	□	□	□	□
It is important to me to understand the work that I do in science class.	□	□	□	□	□
I would prefer to find out why something happens by doing an experiment than by being told.	□	□	□	□	□
Doing experiments is not as good as finding out information from teachers.	□	□	□	□	□
I would prefer to do experiments than to read about them.	□	□	□	□	□
I would prefer to do my own experiments than to find out information from a teacher.	□	□	□	□	□
I would rather find out about things by asking an expert than by doing an experiment.	□	□	□	□	□
It is better to ask the teacher the answer than to find it out by doing experiments.	□	□	□	□	□
I would prefer to do an experiment on a topic than to read about it in science magazines.	□	□	□	□	□
It is better to be told scientific facts than to find them out from experiments.	□	□	□	□	□
I would rather agree with other people than do an experiment to find out for myself.	□	□	□	□	□
I would rather solve a problem by doing an experiment than be told the answer.	□	□	□	□	□

Which of the following courses would you like to take in high school? Please check ALL OF THE BOXES that apply.

□ Astronomy
□ Earth Science
□ Geometry

□ Trigonometry
□ Physics
□ Environmental Science

□ Biology
□ Geobiology
□ Calculus

□ Chemistry
□ Algebra
□ Astrotrigology

	Strongly disagree	Disagree	Unsure	Agree	Strongly agree
The tutors were valuable and gave me good feedback and help.	□	□	□	□	□
The tutors were clearly reacting to the things that we were doing in the software.	□	□	□	□	□
I felt like the tutors knew when we were getting better at investigating the system.	□	□	□	□	□
The tutors did a good job of giving us advice when we were wrong.	□	□	□	□	□
The tutors did a good job of letting us know when we were right about things.	□	□	□	□	□
The tutors were helpful in demonstrating things that we did not really understand.	□	□	□	□	□
Gabriel the Guide was a useful resource to us.	□	□	□	□	□
Craig the Critic was a useful resource to us.	□	□	□	□	□
Marvin the Mentor was a useful resource to us.	□	□	□	□	□
Isla the Interviewer was a useful resource to us.	□	□	□	□	□
I sometimes found myself more interested in how the tutors worked than in the lessons.	□	□	□	□	□
I sometimes found myself looking through the tutors' information without any real question in mind.	□	□	□	□	□
I found the tutors' faces made it more clear what kind of feedback they were giving.	□	□	□	□	□

- Did you have any positive experiences with the tutors used in the software? If so, please describe them.

- Did you have any negative experience with the tutors used in the software? If so, please describe them.

- What would you change about the tutors used in the software? Name as many things as you would like.

- What would you change about the software used as a whole? Name as many things as you would like.

APPENDIX B: CONTENT TEST

The following is the original content test developed for testing this intervention. The objective of this test is to measure students' understanding of ecology, the nature of science, model construction, and the scientific process.

1. A valid hypothesis in any scientific experiment must be:
   a. believable
   b. put in the form of a question
   c. testable
   d. based on past valid experimentation

2. A hypothesis is:
   a. a possible, tentative answer to a question
   b. a theory
   c. an accepted belief that has been tested
   d. a false belief that must be tested

3. A theory is:
   a. a re-creation of an event
   b. an explanation for a phenomena
   c. a plausible, scientifically accepted generalization
   d. a questionable explanation for a natural phenomenon

4. Scientists use scientific models of systems to:
   a. show other scientists their ideas
   b. drive their further research on the system
   c. simulate what might happen to the system in different situations
   d. all of the above

5. Arguing in scientific research is:
   a. Good; it shows which ideas are most popular.
   b. Bad; it means the scientists don't understand the system yet.
   c. Good; it forces scientists to defend their ideas.
   d. Bad; it forces scientists to choose sides.

6. For a scientific model to be good, it must:
   a. Be accepted.
   b. Be supported by evidence.
   c. Be simple.
   d. Be complicated.

7. Which of the following would be bad evidence that pollution is killing insects?
   a. Observations that insect levels drop when pollution levels spike.
   b. Controlled experiments showing insects die in the presence of pollution.
   c. A simulation showing that pollution levels predict insect populations.
   d. An announcement from a TV station that pollution is killing insects.

8. A scientist has found evidence that their hypothesis was wrong. How should they react?
   a. They should be glad; dismissing false ideas is an important part of science.
   b. They should apologize; if they were a better scientist, they

would not have had a wrong hypothesis.
    c. They should stand firm; science is about finding evidence only to support your ideas.
    d. They should hide the evidence; they do not want people knowing their idea was wrong.

9. An example of an abiotic factor in an environment is:
    a. soil/water chemistry
    b. bacteria
    c. plants
    d. other animals

10. An example of a biotic factor in an environment is:
    a. sunlight
    b. plants
    c. oxygen
    d. temperature

11. Which of the following is a part of an ecosystem that can be removed without affecting the ecosystems functionality?
    a. plants
    b. water
    c. sunlight
    d. none of the above

12. Bacteria play a vital role in Earth's ecosystems by
    a. using oxygen
    b. producing acetone
    c. decomposing dead organisms
    d. concentrating valuable minerals

13. Bacteria are essential to life because they _________________ nutrients for living things.
    a. recycle
    b. deposit
    c. digest
    d. engulf

14. Dr. Smith thinks he sees a flaw in Dr. Jordan's model and corrects him. How should Dr. Jordan react?
    a. Accept Dr. Smith's correction and fix his model.
    b. Argue against Dr. Smith's correction about why his model was not incorrect.
    c. Examine Dr. Smith's correction and decide whether it is accurate.
    d. Abandon his model because Dr. Smith is clearly better qualified to model the system.

15. Which of the following does not have to be true about a good scientist?
    a. They look for weaknesses in their own understanding.
    b. They always come to the right answer in the end.
    c. They know how to approach new problems.
    d. They accept correction and feedback from their peers.

APPENDIX C: LESSON PLAN FOR THE CLASSROOM ACTIVITY

The following table provides the lessons that took place during the nine-day intervention.

Georgia state standards addressed by the lesson plan: S7CS2, S7CS3, S7CS4, S7CS5, S7CS6, S7CS7, S7CS8, S7CS9, S7L4, S8P1.

<table>
<tr><th></th><th>Control Classes</th><th>Experimental Classes</th></tr>
<tr><td><b>05/06</b></td><td colspan="2">Pre-Test and Pre-Survey, Introduction to the Lake Clara Meer Problem
No MILA</td></tr>
<tr><td><b>05/07</b></td><td>Lesson 1: Introduction to MILA
With MILA, <b>without</b> tutors</td><td>Lesson 1: Introduction to MILA
With MILA, <b>with</b> tutors</td></tr>
<tr><td><b>05/08</b></td><td colspan="2">Lesson 2: Biological and Ecological Content Knowledge
No MILA</td></tr>
<tr><td><b>05/09</b></td><td>Lesson 3: Introduction to Mechanism
With MILA, <b>without</b> tutors</td><td>Lesson 3: Introduction to Mechanism
With MILA, <b>with</b> tutors</td></tr>
<tr><td><b>05/10</b></td><td>Lesson 4: Introduction to Simulations
With MILA, <b>without</b> tutors</td><td>Lesson 4: Introduction to Simulations
With MILA, <b>with</b> tutors</td></tr>
<tr><td><b>05/13</b></td><td colspan="2">Lesson 5: Lake Water Quality Experiment
No MILA</td></tr>
<tr><td><b>05/14</b></td><td>Lesson 6: Free Modeling
With MILA, <b>without</b> tutors</td><td>Lesson 6: Free Modeling
With MILA, <b>with</b> tutors</td></tr>
<tr><td><b>05/15</b></td><td colspan="2">Lesson 7: Atlanta Temperatures
With MILA, <b>without</b> tutors</td></tr>
<tr><td><b>05/16</b></td><td colspan="2">Post-Test and Post-Survey, Q&A with Georgia Tech Researcher
No MILA</td></tr>
</table>

APPENDIX D: SCRIPTS FOR TEACHER INTERVIEWS

At the conclusion of the study, teacher participants in the intervention were interviewed about their experience. The following is the script used for conducting these interviews.

"Thank you for agreeing to be interviewed about your experiences in this study. Before we begin, I want to give you a couple notes. First, please know that there are no wrong answers to any of these questions. Second, in order to keep the results of this interview useful for analysis, I won't be engaging in any conversation, and I will simply be asking you the questions in this script. Third, please feel free to be critical. Our main goal is to improve the software and unit for future use, and in order to do so, we need to get feedback on what can be improved. So, please don't hesitate to criticize any element of the intervention.

First, do you have any general thoughts on how the study as a whole went?

Did you notice any differences in how students were engaged during the past two weeks compared to how engaged they were in the traditional classroom before?

Did you notice any particular benefits to using MILA itself, separate from the tutors or the lessons and unit as a whole?

Did you notice any particular drawbacks or weaknesses to MILA that interfered with the classroom or confused students?

Are there any particular changes or improvements that you would suggest be made to MILA itself, separate from the tutors or the lessons and unit as a whole?

Did you notice any differences between the students who had the tutors and the students who did not?

Did you see any particular benefits for the classes that had the tutors?

Did you see any particular drawbacks for the classes that had the tutors?

Are there any changes you would make to the tutors for next time?

Do you have any suggestions for any improvements that could be made to the unit as a whole, separate from MILA and the tutors?

Do you have any final thoughts on the intervention?"

REFERENCES

Aleven, V. & Koedinger, K. R. (2002). An effective metacognitive strategy: Learning by doing and explaining with a computer-based Cognitive Tutor. *Cognitive Science, 26*(2), 147-179.

Aleven, V., McLaren, B., Roll, I., & Koedinger, K. R. (2006). Toward metacognitive tutoring: a model of help seeking with a cognitive tutor. *International Journal of Artificial Intelligence and Education*, *16*(2), 101-128.

Aleven, V., McLaren, B. M., Sewall, J., & Koedinger, K. R. (2009). A new paradigm for intelligent tutoring systems: Example-tracing tutors. *International Journal of Artificial Intelligence in Education*, *19*(2), 105-154.

Anderson, J., Corbett, A., Koedinger, K., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *The Journal of Learning Sciences, 4*, 167-207.

Anderson, J. R., Reder, L. M., & Simon, H. A. (1996). Situated learning and education. *Educational Researcher, 25*(4), 5-11.

Azevedo, R., Cromley, J.G., Moos, D.C., Greene, J.A., Winters, F.I. (2011). Adaptive content and process scaffolding: A key to facilitating students' self-regulated learning with hypermedia. *Psychological Testing and Assessment Modeling, 53*, 106-140.

Azevedo, R., Witherspoon, A., Chauncey, A., Burkett, C., & Fike, A. (2009). MetaTutor: A MetaCognitive tool for enhancing self-regulated learning. In R. Pirrone, R. Azevedo, & G. Biswas (Eds.), *Proceedings of the AAAI Fall Symposium on Cognitive and Metacognitive Educational Systems*, 14-19.

Babbs, P. J., & Moe, A. J. (1983). Metacognition: A key for independent learning from text. *The Reading Teacher*, 422-426.

Baker, L., & Brown, A. L. (1984). Metacognitive skills and reading. *Handbook of Reading Research*, *1*, 353-394.

Baker, R., Corbett, A., Roll, I., & Koedinger, K. (2008). Developing a generalizable detector of when students game the system. *User Modeling and User-Adapted Interaction 18*(3), 287-314.

Barica, J., & Mathias, J. A. (1979). Oxygen depletion and winterkill risk in small prairie lakes under extended ice cover. *Journal of the Fisheries Board of Canada*, *36*(8), 980-986.

Bencze, L., & Hodson, D. (1999). Changing practice by changing practice: Toward more authentic science and science curriculum development. *Journal of Research in Science Teaching*, *36*(5), 521-539.

Berland, L. K., & Reiser, B. J. (2009). Making sense of argumentation and explanation. *Science Education*, *93*(1), 26-55.

Biswas, G., Leelawong, K., Schwartz, D., & Vye, N. (2005). Learning By Teaching: A New Agent Paradigm For Educational Software. *Applied Artificial Intelligence 19*(3-4). 363-392.

Biswas, G., Kinnebrew, J., & Segedy, J. (2013). Analyzing Students' Metacognitive Strategies in Open-Ended Learning Environments. In *Proc. of the 35th Annual Meeting of the Cognitive Science Society*, 209-214. Berlin.

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.

Borkowski, J. G., Carr, M., Rellinger, E., & Pressley, M. (1990). Self-regulated cognition: Interdependence of metacognition, attributions, and self-esteem. *Dimensions of Thinking and Cognitive Instruction, 1*, 53-92.

Borkowski, J., Chan, L., & Muthukrishna, N. (2000). A Process-Oriented Model of Metacognition: Links Between Motivation and Executive Functioning. In Schraw, G. & Impara, J. (Eds.) *Issues in Measurement of Metacognition*, 1-41

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

Braund, M., & Reiss, M. (2006). Towards a more authentic science curriculum: The contribution of out-of-school learning. *International Journal of Science Education*, *28*(12), 1373-1388.

Bray, J. N. (Ed.). (2000). *Collaborative inquiry in practice: Action, reflection, and making meaning*. Sage.

Brown, A. (1987). Metacognition, executive control, self-regulation, and other more mysterious mechanisms. In F. Reiner, & R. Kluwe (Eds.), *Metacognition, Motivation, and Understanding*, 65-116. Hillsdale, NJ: Erlbaum.

Bruner, J. S. (1961). The act of discovery. *Harvard Educational Review, 31*(1), 21-32.

Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research*, *65*(3), 245-281.

Clement, J. (2000). Model-based learning as a key research area for science education. *International Journal of Science Education*, *22*(9), 1041-1053.

Clement, J. (2008). *Creative Model Construction in Scientists and Students: The Role of Imagery, Analogy, and Mental Simulation*. Dordrecht: Springer.

Cohn, J. P. (2008). Citizen science: Can volunteers do real research? *BioScience*, *58*(3), 192-197.

Cook, D. L. (1962). The Hawthorne effect in educational research. *Phi Delta Kappan*, 116-122.

Corbett, A. (2001). Cognitive computer tutors: Solving the two-sigma problem. In *User Modeling 2001*, 137-147. Berlin: Springer.

Crawford, S., & Stucki, L. (1990). Peer review and the changing research record. *Journal of the American Society for Information Science, 41*(3), 223-228.

Crosby, M. E., & Iding, M. K. (1997). The influence of a multimedia physics tutor and user differences on the development of scientific knowledge. *Computers & Education*, *29*(2), 127-136.

D'Mello, S., & Graesser, A. (2013). Design of Dialog-Based Intelligent Tutoring Systems to Simulate Human-to-Human Tutoring. In *Where Humans Meet Machines*, 233-269. New York: Springer.

Darden, L. (1998) Anomaly-Driven Theory Redesign: Computational Philosophy of Science Experiments. In T. Bynum & J. Moor (Eds.), *The Digital Phoenix: How Computers are Changing Philosophy*, 62-78. New York: Blackwell Publishers.

Davies, J., Nersessian, N., & Goel, A. (2005). Transfer in Visual Case-Based Problem-Solving. In H. Munoz-Avila & F. Ricci (Eds.), *Proceedings of the 6th International Conference on Case-Based Reasoning*, 163-176. Chicago.

Dweck, C. (2000). *Self-theories: Their role in motivation, personality, and development*. Philadelphia, PA: Psychology Press.

Dweck, C. (2006). *Mindset: The New Psychology of Success*. Random House.

Edelson, D. (1997). Realizing authentic scientific learning through the adaptation of scientific practice. In K. Tobin & B. Fraser (Eds.), *International Handbook of Science Education*. Dordrecht, NL: Kluwer.

Fraser, B. J. (1981). *Test of science related attitudes.* Australian Council for Educational Research. Hawthorn, Victoria.

Gama, C. (2004). Metacognition in interactive learning environments: The reflection assistant model. In *7th Conference on Intelligent Tutoring Systems*, 668–677. Berlin: Springer.

Ganz, M. N., & Ganz, B. C. (1990). Linking metacognition to classroom success. *The High School Journal*, 180-185.

Ghahramani, Z. (2001). An introduction to hidden Markov models and Bayesian networks. *International Journal of Pattern Recognition and Artificial Intelligence*, *15*(1), 9-42.

Gibson, H. L., & Chase, C. (2002). Longitudinal impact of an inquiry-based science program on middle school students' attitudes toward science. *Science Education*, *86*(5), 693-705.

Gilbert, J. K. (2004). Models and modelling: Routes to more authentic science education. *International Journal of Science and Mathematics Education*, *2*(2), 115-130.

Gilbert, J. K., Boulter, C. J., & Elmer, R. (2000). Positioning models in science education and in design and technology education. In *Developing models in science education* (pp. 3-17). Springer Netherlands.

Gobert, J. D., Sao Pedro, M. A., Baker, R. S., Toto, E., & Montalvo, O. (2012). Leveraging educational data mining for real-time performance assessment of scientific inquiry skills within microworlds. *Journal of Educational Data Mining*, *4*(1), 111-143.

Goel, A. K., Bras, B., Helms, M., Rugaber, S., Tovey, C., Vattam, S., Weissburg, M., Wiltgen, B., & Yen, J. (2011). Design patterns and cross-domain analogies in biologically inspired sustainable design. In *2011 AAAI Spring Symposium Series*.

Goel, A. & Joyner, D. (2014). Computational Ideation in Scientific Discovery: Interactive Construction, Evaluation, and Revision of Conceptual Models. In *Workshops at the Twenty-Eighth AAAI Conference on Artificial Intelligence.* Quebec City, Quebec.

Goel, A., Rugaber, S., Joyner, D. A., Vattam, S., Hmelo-Silver, C., Jordan, R., Sinha, S., Honwad, S., & Eberbach, C. (2013). Learning Functional Models of Aquaria: The ACT Project on Ecosystem Learning in Middle School Science. In R. Azevedo & V. Aleven (Eds.) *International Handbook on Meta-Cognition and Self-Regulated Learning*, 545-559. New York: Springer.

Goel, A. K., Rugaber, S., & Vattam, S. (2009). Structure, behavior, and function of complex systems: The structure, behavior, and function modeling language.

Artificial Intelligence for Engineering Design, Analysis and Manufacturing, 23(1), 23-35.

Goel, A. K., Vattam, S., Rugaber, S., Joyner, D. A., Hmelo-Silver, C., Jordan, R., Honwad, S., Gray, S., & Sinha, S. (2010). Learning Functional and Causal Abstractions of Classroom Aquaria. In *Proceedings of the 32nd Annual Meeting of the Cognitive Science Society*. Portland, Oregon.

Goel, A. K., Vattam, S., Helms, M., & Wiltgen, B. (2011). An information-processing account of creative analogies in biologically inspired design. In *Proceedings of the 8th ACM conference on Creativity and Cognition*, 71-80. ACM.

Goldin, I., Renken, M., Galyardt, A., & Litkowski, E. (2014). Individual Differences in Identifying Sources of Science Knowledge. In *Open Learning and Teaching in Educational Communities*, 152-164. Springer International Publishing.

Graham, E. A., Henderson, S., & Schloss, A. (2011). Using mobile phones to engage citizen scientists in research. *Eos, Transactions American Geophysical Union*, *92*(38), 313-315.

Grasha, A. (1996). *Teaching with Style*. Pittsburgh: Alliance Publishers.

Griffith, T., Nersessian, N., & Goel, A. (2000). Function-follows-Form: Generative Modeling in Scientific Reasoning. In *Proceeds of the 22nd Cognitive Science Conference*, 196-201.

Guzdial, M., & Adams, J. C. (2014). MOOCs need more work; so do CS graduates. *Commun. ACM*, *57*(1), 18-19.

Heffernan, N. T., & Koedinger, K. R. (2002). An intelligent tutoring system incorporating a model of an experienced human tutor. In *Intelligent Tutoring Systems* (pp. 596-608). Springer Berlin Heidelberg.

Helms, M., Vattam, S., Goel, A. K., & Yen, J. (2011). Enhanced Understand of Biological Systems Using Structure-Behavior-Function Models. In *Proc. of 11th*

IEEE International Conference on Advanced Learning Technologies, 239-243. IEEE.

Hollan, J., Hutchins, E., & Kirsh, D. (2000). Distributed cognition: toward a new foundation for human-computer interaction research. *ACM Transactions on Computer-Human Interaction*, *7*(2), 174-196.

Honwad, S., Hmelo-Silver, C., Jordan, R., Eberbach, C., Gray, S., Sinha, S., Goel, A. K., Vattam, S., Rugaber, S., & Joyner, D. A. (2010). Connecting the Visible to the Invisible: Helping Middle School Students Understand Complex Ecosystem Processes. In *Proceedings of the 32nd Annual Meeting of the Cognitive Science Society*, 133-138. Portland, Oregon.

Howard, E., & Davis, A. K. (2009). The fall migration flyways of monarch butterflies in eastern North America revealed by citizen scientists. *Journal of Insect Conservation*, *13*(3), 279-286.

Isaacson, R. M., & Fujita, F. (2006). Metacognitive Knowledge Monitoring and Self-Regulated Learning: Academic Success and Reflections on Learning. *Journal of Scholarship of Teaching and Learning*, *6*(1), 39-55.

Jackson, S. L. (1995). The ScienceWare Modeler: a learner-centered tool for students building models. In *Conference Companion on Human Factors in Computing Systems*, 7-8. ACM.

Johnson, W., Rickel, J., & Lester, J. (2000). Animated pedagogical agents: Face-to-face interaction in interactive learning environments. *International Journal of Artificial Intelligence in Education*, *11*, 47-78.

Joyner, D. A., & Goel, A. (2014). Attitudinal Gains from Engagement with Metacognitive Tutors in an Exploratory Learning Environment. In *Proceedings of the 12th International Conference on Intelligent Tutoring Systems*, 627-628. Honolulu, Hawaii.

Joyner, D. A., & Goel, A. (2015). Improving Inquiry-Driven Modeling in Science Education through Interaction with Intelligent Tutoring Agents. In *Proceedings of the 20th International Conference on Intelligent User Interfaces.* Atlanta, Georgia.

Joyner, D. A., Goel, A., & Papin, N. (2014). MILA–S: Generation of agent-based simulations from conceptual models. In *Proceedings of the 19th International Conference on Intelligent User Interfaces,* 289-298. Haifa, Israel.

Joyner, D. A., Goel, A., Rugaber, S., Hmelo-Silver, C., & Jordan, R. (2011). Evolution of an Integrated Technology for Supporting Learning about Complex Systems: Looking Back, Looking Ahead. In *Proc. Of 11th International Conference on Advanced Learning Technologies,* 257-259. Athens, Georgia.

Joyner, D. A., Goel, A., & Majerich, D. (2013). Metacognitive Tutoring for Scientific Modeling. In *AIED 2013 Workshop Proceedings: Scaffolding in Open-Ended Learning Environments,* 37-40. Memphis, Tennessee.

Joyner, D. A., Majerich, D., & Goel, A. (2013). Facilitating authentic reasoning about complex systems in middle school science education. In *Proc. of 11th Annual Conference on Systems Engineering Research*, 1043-1052. Atlanta.

Kemeny, J. G., & Snell, J. L. (1960). *Finite Markov chains* (Vol. 356). Princeton, NJ: van Nostrand.

Ketelhut, D. J. (2007). The impact of student self-efficacy on scientific inquiry skills: An exploratory investigation in River City, a multi-user virtual environment. *Journal of Science Education and Technology, 16*(1), 99-111.

Keys, C. W., & Bryan, L. A. (2001). Co-constructing inquiry-based science with teachers: Essential research for lasting reform. *Journal of Research in Science Teaching*, *38*(6), 631-645.

Kinnebrew, J., Biswas, G., Sulcer, B., & Taylor, R. (2012). Investigating Self-Regulated Learning in Teachable Agent Environments. In R. Azevedo & V. Aleven (Eds.), *International Handbook of Metacognition and Learning Technologies: Vol. 99. Springer International Handbooks of Education*. New York: Springer.

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist, 41*(2), 75-86.

Klahr, D., & Nigam, M. (2004). The equivalence of learning paths in early science instruction effects of direct instruction and discovery learning. *Psychological Science*, *15*(10), 661-667.

Knox, J., Bayne, S., MacLeod, H., Ross, J., & Sinclair, C. (2012). MOOC Pedagogy: the challenges of developing for Coursera. *ALT Online Newsletter*, *28*(08.08), 2012.

Lajoie, S., Lavigne, N., Guerrera, C., & Munsie, S. (2001). Constructing knowledge in the context of Bio World. *Instructional Science, 29,* 155-186.

Landine, J., & Stewart, J. (1998). Relationship between Metacognition, Motivation, Locus of Control, Self-Efficacy, and Academic Achievement. *Canadian Journal of Counselling, 32*(3), 200-212.

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Cambridge University Press.

Lee, H. S., & Butler, N. (2003). Making authentic science accessible to students. *International Journal of Science Education*, *25*(8), 923-948.

Lee, M. J., & Ko, A. J. (2011). Personifying programming tool feedback improves novice programmers' learning. In *Proceedings of the 7th International Conference on Computing Education Research*, 109-116. ACM.

Lehrer, R., & Schauble, L. (2004). Modeling natural variation through distribution. *American Educational Research Journal, 41*(3), 635-679.

Lepper, M. R., & Chabay, R. W. (1988). Socializing the intelligent tutor: Bringing empathy to computer tutors. In *Learning Issues for Intelligent Tutoring Systems*, 242-257. Springer.

Lewin, T. (2013). Master's degree is new frontier of study online. *New York Times*.

Lintean, M., Rus, V., & Azevedo, R. (2012). Automatic Detection of Student Mental Models during Prior Knowledge Activation in MetaTutor. *International Journal of Artificial Intelligence in Education*, *21*(3), 169-190.

Löhner, S., van Joolingen, W. R., Savelsbergh, E. R., & van Hout-Wolters, B. (2005). Students' reasoning during modeling in an inquiry learning environment. *Computers in Human Behavior*, *21*(3), 441-461.

López, G. G. I., Hermanns, H., & Katoen, J. P. (2001). Beyond memoryless distributions: Model checking semi-Markov chains. In *Process Algebra and Probabilistic Methods. Performance Modelling and Verification*, 57-70. Berlin: Springer.

Lovett, M. C. (2008). Teaching metacognition. In Presentation to the Educause Learning Initiative Annual Meeting, 29 January 2008.

Luther, K., Counts, S., Stecher, K. B., Hoff, A., & Johns, P. (2009). Pathfinder: an online collaboration environment for citizen scientists. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 239-248. ACM.

MacIsaac, D., & Falconer, K. (2002). Reforming physics instruction via RTOP. *The Physics Teacher*, *40*(8), 479-485.

Manlove, S., Lazonder, A. W., & Jong, T. D. (2006). Regulative support for collaborative scientific inquiry learning. *Journal of Computer Assisted Learning*, *22*(2), 87-98.

Martin, F. (2013). Fight the MOOC-opalypse! and Reflections on the aporia of learning. *Journal of Computing Sciences in Colleges*, *28*(6), 5-6.

Mathan, S. A., & Koedinger, K. R. (2005). Fostering the intelligent novice: Learning from errors with metacognitive tutoring. *Educational Psychologist*, *40*(4). 257-265.

Matsuda, N., & VanLehn, K. (2005). Advanced Geometry Tutor: An intelligent tutor that teaches proof-writing with construction. In *Proceedings of The 12th International Conference on Artificial Intelligence in Education*, 443-450. Amsterdam: IOS Press.

Mayer, R. E. (2004). Should there be a three-strikes rule against pure discovery learning? *American Psychologist*, *59*(1), 14-19.

Merrill, D. C., Reiser, B. J., Ranney, M., & Trafton, J. G. (1992). Effective tutoring techniques: A comparison of human tutors and intelligent tutoring systems. *The Journal of the Learning Sciences*, *2*(3), 277-305.

National Research Council. (1996). *National Science Education Standards.* Washington, DC: National Academy Press.

Nersessian, N. (1992). How do scientists think? Capturing the dynamics of Conceptual change in science. In R. Giere (Ed.), *Minnesota Studies in the Philosophy of Science Vol. XV: Cognitive Models of Science*. Minneapolis: University of Minnesota Press.

Nersessian, N. (1999). Model-based reasoning in conceptual change. In L. Magnani, N. Nersessian, & P. Thagard (Eds.), *Model-Based Reasoning in Scientific Discovery*. New York: Kluwer/Plenum Publishers.

Nersessian, N. (2008). *Creating Scientific Concepts*. Cambridge, MA: MIT Press.

Olney, A., D'Mello, S., Person, N., Cade, W., Hayes, P., Williams, C., Lehman, B., & Graesser, A. (2012). Guru: A Computer Tutor that Models Expert Human Tutors. In *Proc. of the 11th International Conference on Intelligent Tutoring Systems*, 256-261. Berlin: Springer.

Oscarson, D. B., & Calhoun, A. J. (2007). Developing vernal pool conservation plans at the local level using citizen-scientists. *Wetlands, 27*(1), 80-95.

Palinscar, A. S., & Brown, A. L. (1984). Reciprocal teaching of comprehension-fostering and comprehension-monitoring activities. *Cognition and Instruction, 1*(2), 117-175.

Pappano, L. (2012). The Year of the MOOC. *The New York Times, 2*(12), 2012.

Prabhakar, S., & Goel, A. K. (1998). Functional modeling for enabling adaptive design of devices for new environments. *Artificial intelligence in Engineering, 12*(4), 417-444.

Peffer, M. (2014). Choose Your Own Science Adventure [software]. Available from http://www.cyosa.us/

Razzouk, R. & Shute, V. J. (2012). What is design thinking and why is it important?. *Review of Educational Research, 82*(3), 330-348.

Reddy, S., Chen, G., Fulkerson, B., Kim, S. J., Park, U., Yau, N., Cho, J., Hansen, M., & Heidemann, J. (2007). Sensor-Internet share and search: Enabling collaboration of citizen scientists. In *Proceedings of the ACM Workshop on Data Sharing and Interoperability on the World-wide Sensor Web*, 11-16.

Renken, M. D., & Nunez, N. (2013). Computer simulations and clear observations do not guarantee conceptual understanding. *Learning and Instruction*, *23*, 10-23.

Renken, M., Carrion, C., & Litkowski, E. (2014). Targeting Students' Epistemologies: Instructional and Assessment Challenges to Inquiry-Based Science Education. *Inquiry-based Learning for Faculty and Institutional Development: A Conceptual and Practical Resource for Educators (Innovations in Higher Education, Teaching and Learning, 1*, 147-174. Emerald Group Publishing.

Ritter, S., Anderson, J. R., Koedinger, K. R., & Corbett, A. (2007). Cognitive Tutor: Applied research in mathematics education. *Psychonomic Bulletin & Review, 14*(2), 249-255.

Roll, I., Aleven, V., McLaren, B., & Koedinger, K. (2007). Designing for metacognition—applying cognitive tutor principles to the tutoring of help seeking. *Metacognition in Learning 2*(2), 125-140.

Roll, I., Aleven, V., & Koedinger, K. R. (2010). The invention lab: Using a hybrid of model tracing and constraint-based modeling to offer intelligent support in inquiry environments. In V. Aleven, J. Kay, & J. Mostow (Eds.), In *Proceedings of the international conference on intelligent tutoring systems*, 115-24. Berlin: Springer.

Roll, I., Aleven, V., McLaren, B. M., & Koedinger, K. R. (2011). Metacognitive practice makes perfect: Improving students' self-assessment skills with an intelligent tutoring system. In *Proceedings of the 15th International Conference on Artificial Intelligence in Education*, 288-295. Berlin: Springer.

Roll, I., Holmes, N., Day, J., & Bonn, D. (2012). Evaluating metacognitive scaffolding in guided invention activities. *Instructional Science, 40*(4), 691-710. Springer.

Ross, J., Sinclair, C., Knox, J., Bayne, S., & Macleod, H. (2014). Teacher experiences and academic identity: The missing components of MOOC pedagogy. *MERLOT Journal of Online Learning and Teaching*, *10*(1), 56-68.

Rudman, P., & Zajicek, M. (2006). Autonomous Agent as Helper-Helpful or Annoying? In *International Conference on Intelligent Agent Technology 2006*, 170-176. IEEE.

Schiaffino, S., & Amandi, A. (2004). User–interface agent interaction: personalization issues. *International Journal of Human-Computer Studies*, *60*(1), 129-148.

Schraw, G., Crippen, K. J., & Hartley, K. (2006). Promoting self-regulation in science education: Metacognition as part of a broader perspective on learning. *Research in Science Education*, *36*(1-2), 111-139.

Schwarz, C. V., Reiser, B. J., Davis, E. A., Kenyon, L., Achér, A., Fortus, D., Shwartz, Y., Hug, B. & Krajcik, J. (2009). Developing a learning progression for scientific modeling: Making scientific modeling accessible and meaningful for learners. *Journal of Research in Science Teaching, 46*(6), 632-654.

Schwarz, C. V., & White, B. Y. (2005). Metamodeling knowledge: Developing students' understanding of scientific modeling. *Cognition and Instruction*, *23*(2), 165-205.

Schweingruber, H. A., Duschl, R. A., & Shouse, A. W. (Eds.). (2007). *Taking Science to School: Learning and Teaching Science in Grades K-8*. National Academies Press.

Schiff, S. K. (1970). Research for Tomorrow's Schools: Disciplined Inquiry for Education. *Archives of General Psychiatry*, *22*(4), 381.

Shaffer, D. W., & Resnick, M. (1999). "Thick" Authenticity: New Media and Authentic Learning. *Journal of Interactive Learning Research*, *10*(2), 195-215.

Shulman, L. & Keisler, E. (Eds.) (1966). *Learning by Discovery: A critical appraisal.* Chicago: Rand McNally.

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1),153-189.

Silvertown, J. (2009). A new dawn for citizen science. *Trends in Ecology & Evolution*, *24*(9), 467-471.

Sinha, S., Gray, S., Hmelo-Silver, C., Jordan, R., Honwad, S., Eberbach, C., Rugaber, S., Vattam, S., & Goel, A. K. (2010). Appropriating Conceptual Representations: A Case of Transfer in a Middle School Science Teacher. In *Proceedings of the Ninth International Conference of the Learning Sciences*, 834-841. Chicago.

Sins, P. H., Savelsbergh, E. R., & van Joolingen, W. R. (2005). The Difficult Process of Scientific Modelling: An analysis of novices' reasoning during computer-based modelling. *International Journal of Science Education*, *27*(14), 1695-1721.

Soloway, E., Pryor, A. Z., Krajcik, J. S., Jackson, S., Stratford, S. J., Wisnudel, M., & Klein, J. T. (1997). ScienceWare's Model-It: Technology to Support Authentic Science Inquiry. *Technological Horizons in Education*, *25*(3), 54-56.

Spiro, R., Feltovich, R., Jacobson, M., & Coulson, R. (1992). Cognitive flexibility, constructivism, and hypertext: Random access instruction for advanced knowledge acquisition in ill-structured domains. In T. M. Duffy & D. H. Jonassen (Eds.), *Constructivism and the Technology of Instruction: A Conversation*. 57-76. Hillsdale, NJ: Lawrence Erlbaum Associates.

Stapleton, C., Smith, E., & Hughes, C. E. (2005,). The art of nurturing citizen scientists through mixed reality. In *Proceedings of the 4th IEEE/ACM International Symposium on Mixed and Augmented Reality*, 2-11. IEEE Computer Society.

Sternberg, R. J. (1998). Metacognition, abilities, and developing expertise: What makes an expert student? *Instructional Science, 26*(1-2), 127-140.

Stratford, S. J., & Finkel, E. A. (1996). The impact of scienceware and foundations on students' attitudes towards science and science classes. *Journal of Science Education and Technology*, *5*(1), 59-67.

Suthers, D. D., Toth, E. E., & Weiner, A. (1997). An integrated approach to implementing collaborative inquiry in the classroom. In *Proc. of the 2nd International Conference on Computer Support for Collaborative Learning*, 275-282.

Svoboda, J., & Passmore, C. (2013). The strategies of modeling in biology education. *Science & Education, 22*(1), 119-142.

Ting, C. Y., Zadeh, M. R. B., & Chong, Y. K. (2006). A decision-theoretic approach to scientific inquiry exploratory learning environment. In *Proc. of the 8th International Conference on Intelligent Tutoring Systems*, 85-94. Berlin: Springer.

Tinker, R. and Wilensky, U. (2007). NetLogo Climate Change model. http://ccl.northwestern.edu/netlogo/models/ClimateChange. Center for Connected Learning and Computer-Based Modeling, Northwestern Institute on Complex Systems, Northwestern University, Evanston, IL.

Townsend, S. A., Boland, K. T., & Wrigley, T. J. (1992). Factors contributing to a fish kill in the Australian wet/dry tropics. *Water Research*, *26*(8), 1039-1044.

van Joolingen, W. R., de Jong, T., Lazonder, A. W., Savelsbergh, E. R., & Manlove, S. (2005). Co-Lab: Research and development of an online learning environment for collaborative scientific discovery learning. *Computers in Human Behavior*, *21*(4), 671-688.

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

Vattam, S., Goel, A. K., & Rugaber, S. (2011). Behavior Patterns: Bridging Conceptual Models and Agent-Based Simulations in Interactive Learning Environments. In *Proceedings of the 11th IEEE International Conference on Advanced Learning Technologies*, 139-141. Athens, Georgia.

Vattam, S., Goel, A. K., Rugaber, S., Hmelo-Silver, C., Jordan, R., Gray, S., & Sinha, S. (2011A). Understanding Complex Natural Systems by Articulating Structure-Behavior-Function Models. *Educational Technology & Society, Special Issue on Creative Design, 14*(1), pp. 66-81.

Vattam, S., Wiltgen, B., Helms, M., Goel, A. K., & Yen, J. (2011B). DANE: fostering creativity in and through biologically inspired design. In *Design Creativity 2010*, 115-122. London: Springer.

Waller, P. (2014). Encyclopedia of Life. *The American Biology Teacher*, *76*(6), 420-421.

Weinburgh, M.E. & Steele, D. (2000). The modified attitudes toward science inventory: Developing an instrument to be used with fifth grade urban students. *Journal of Women and Minorities in Science and Engineering, 6,* 87-94.

Weinert, F. E. (1987). Introduction and overview: Metacognition and motivation as determinants of effective learning and understanding. In Weinert, F. & Kluwe, R. (Eds.) *Metacognition, Motivation, and Understanding*, 1-16.

Wellman, B., & Lipton, L. (2004). *Data-driven dialogue: A facilitator's guide to collaborative inquiry*. Mira Via, LLC.

Wheatley, T. (2009). "Hundreds of fish die at Piedmont Park's Lake Clara Meer." *Creative Loafing*. Retrieved from http://clatl.com/freshloaf/archives/2009/08/10/ hundreds-of-fish-die-at-piedmont-parks-lake-clara-meer-update

White, B. & Frederiksen, J. (1998). Inquiry, modeling, and metacognition: Making science accessible to all students. *Cognition and Instruction, 16*(1). 3–118.

Wilensky, U. (1999). NetLogo. http://ccl.northwestern.edu/netlogo/. Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL.

Wilensky, U. (2005). NetLogo Algae model. http://ccl.northwestern.edu/netlogo/ models/Algae. Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL.

Wiltgen, B., Vattam, S., Helms, M., Goel, A. K., & Yen, J. (2011). Learning Functional Models of Biological Systems for Biologically Inspired Design. In *Proceedings of the 11th IEEE International Conference on Advanced Learning Technologies*, 355-357. IEEE.

Yen, J., Weissburg, M., Helms, M., & Goel, A. (2011). Biologically inspired design: a tool for interdisciplinary education. In Bar-Cohen, Y. (Ed.) *Biomimetics: Nature-Based Innovation*, 331-360. Boca Raton: Taylor & Francis.

CURRICULUM VITAE

Personal Information

David A. Joyner

Office Address: 85 Fifth Street NW, Atlanta, GA 30332

Telephone: 404-429-2380

E-Mail: david.joyner@gatech.edu

Web Site: www.davidjoyner.net

Education

- Ph.D. in Human-Centered Computing, Georgia Tech, 2009-2015

  Specializing in Learning Sciences & Technology
- M.S. in Human-Computer Interaction, Georgia Tech, 2008-2009
- B.S. in Computer Science, Georgia Tech, 2005-2008.

  Certificate in Social & Personality Psychology

  Specializing in People & Media

Research Experience

- **Georgia Institute of Technology**, August 2009 – Present
  - "Fostering Complex Systems Learning in Early Science Education"
  - Supervised by Principal Investigator Dr. Ashok Goel.
- **Georgia Institute of Technology**, August 2008 – May 2009
  - Created, analyzed and tested Java version of standardized Introduction to Computer Science assessment test.
  - Supervised by Investigators Allison Tew and Dr. Mark Guzdial.

- **Georgia Institute of Technology**, January 2008 – April 2008
    - “Investigating the Effects of Variously Moded Educational Materials on Learning”
    - Supervised by Principal Investigator Dr. Jim Foley.

Non-Academic Experience

- Udacity, Mountain View, CA; Course Developer; February 2014 – Present.
- QOil, Atlanta, GA; User Interface Designer; June 2008 – December 2008.
- Media Technology Solutions; Norcross, GA; User Interface Designer; May 2006 – May 2008.
- Self-Employed; Atlanta, GA; Web Developer; January 2005 – August 2009.
- Self-Employed, Atlanta, GA; Private Tutor; August 2003 – Present.

Publications

Book Chapters

- Goel, A., Rugaber, S., **Joyner, D. A.**, Vattam, S., Hmelo-Silver, C., Jordan, R., Sinha, S., Honwad, S., & Eberbach, C. (2013). Learning Functional Models of Aquaria: The ACT Project on Ecosystem Learning in Middle School Science. In R. Azevedo & V. Aleven (Eds.) *International Handbook on Meta-Cognition and Self-Regulated Learning*, 545-559. New York: Springer.

Journal Articles

- Wu, A., **Joyner, D. A.,** & Do, E. Y. (2010). Move, Beam, and Check! Imagineering Tangible Optical Chess on An Interactive Tabletop Display. *ACM Computers in Entertainment 8*(3). (ACE 2010 Best paper award)

Conference Papers

- **Joyner, D. A.** & Goel, A. (2015). Improving Inquiry-Driven Modeling in Science Education through Interaction with Intelligent Tutoring Agents. In *Proceedings of the 20th International Conference on Intelligent User Interfaces.* Atlanta, Georgia.
- **Joyner, D. A.** & Goel, A. (2014). Attitudinal Gains from Engagement with Metacognitive Tutors in an Exploratory Learning Environment. In *Proceedings of the 12th International Conference on Intelligent Tutoring Systems*, 627-628. Honolulu, Hawaii.
- **Joyner, D. A.**, Goel, A., & Papin, N. (2014). MILA–S: Generation of agent-based simulations from conceptual models. In *Proceedings of the 19th International Conference on Intelligent User Interfaces*, 289-298. Haifa, Israel.
- **Joyner, D. A.**, Majerich, D., & Goel, A. (2013). Facilitating authentic reasoning about complex systems in middle school science education. In *Proc. of 11th Annual Conference on Systems Engineering Research*, 1043-1052. Atlanta.
- **Joyner, D. A.**, Goel, A., Rugaber, S., Hmelo-Silver, C., & Jordan, R. (2011). Evolution of an Integrated Technology for Supporting Learning about Complex Systems: Looking Back, Looking Ahead. In *Proc. Of 11th International Conference on Advanced Learning Technologies*, 257-259. Athens, Georgia.
- Hmelo-Silver, C., Jordan, R., Honwad, S., Eberbach, C., Sinha, S., Goel, A., Rugaber, S., & **Joyner, D. A.** (2011). Foregrounding Behaviors and Functions to Promote Ecosystem Understanding. In *Proc. Ninth Hawaii International Conference on Education*.
- Goel, A., Vattam, S., Rugaber, S., **Joyner, D. A.,** Hmelo-Silver, C., Jordan, R., Honwad, S., Gray, S., & Sinha, S. (2010). Learning Functional and Causal Abstractions of Complex Systems. In *Proc. of the 32nd Annual Meeting of the Cogntive Science Society*. Portland, Oregon.

- Honwad, S., Hmelo-Silver, C., Jordan, R., Eberbach, C., Gray, S., Sinha, S., Goel, A. K., Vattam, S., Rugaber, S., & **Joyner, D. A.** (2010). Connecting the Visible to the Invisible: Helping Middle School Students Understand Complex Ecosystem Processes. In *Proceedings of the 32nd Annual Meeting of the Cognitive Science Society*, 133-138. Portland, Oregon.
- **Joyner, D. A.,** Wu, C. & Do, E. Y. Tangible optical chess: a laser strategy game on an interactive tabletop. In *Proc. of the 8th international Conference on Interaction Design and Children*, Como, 278-279. ACM Press.

Workshop Papers

- Goel, A. & **Joyner, D. A.** (2014). Computational Ideation in Scientific Discovery: Interactive Construction, Evaluation, and Revision of Conceptual Models. In *Workshops at the Twenty-Eighth AAAI Conference on Artificial Intelligence.* Quebec City, Quebec.
- Goel, A. K., Kunda, M., **Joyner, D. A.**, & Vattam, S. (2013). Learning about Representational Modality: Design and Programming Projects for Knowledge-Based AI. In *Fourth AAAI Symposium on Educational Advances in Artificial Intelligence*, 1586-1591.
- **Joyner, D. A.**, Goel, A., & Majerich, D. (2013). Metacognitive Tutoring for Scientific Modeling. In *AIED 2013 Workshop Proceedings: Scaffolding in Open-Ended Learning Environments*, 37-40. Memphis, Tennessee.