# Improving Inquiry-Driven Modeling in Science Education through Interaction with Intelligent Tutoring Agents

**David A. Joyner**
Design & Intelligence Lab.
Georgia Institute of Technology
85 5th Street NW
Atlanta, GA 30332
david.joyner@gatech.edu

**Ashok K. Goel**
Design & Intelligence Lab.
Georgia Institute of Technology
85 5th Street NW
Atlanta, GA 30332
goel@cc.gatech.edu

## ABSTRACT
This paper presents the design and evaluation of a set of intelligent tutoring agents constructed to teach teams of students an authentic process of inquiry-driven modeling. The paper first presents the theoretical grounding for inquiry-driven modeling as both a teaching strategy and a learning goal, and then presents the need for guided instruction to improve learning of this skill. However, guided instruction is difficulty to provide in a one-to-many classroom environment, and thus, this paper makes the case that interaction with a metacognitive tutoring system can help students acquire the skill. The paper then describes the design of an exploratory learning environment, the Modeling and Inquiry Learning Application (MILA), and an accompanying set of metacognitive tutors (MILA–T). These tools were used in a controlled experiment with 84 teams (237 total students) in which some teams received and interacted with the tutoring system while other teams did not. The effect of this experiment on teams' demonstration of inquiry-driven modeling are presented.

### Author Keywords
Metacognitive tutoring; scientific discovery; scientific inquiry; discovery-based learning; inquiry-based learning; constructionism; guided instruction.

### ACM Classification Keywords
K.3.1: Computer Uses in Education (Computer-Assisted Instruction); I.2.11: Distributed Artificial Intelligence (Intelligent Agents); J.3 Life and Medical Sciences.

## INTRODUCTION
Constructionism has been one of the dominant theories of modern instructional design for several decades. Originating from the theories of Jean Piaget [38] and Seymour Papert [37], constructionist learning approaches advocate learning in open-ended environments where the learner plays a significant role in driving the learning goals and outcomes. Constructionist learning approaches come in various forms, such as discovery-based learning [19], problem-based learning [14], project-based learning [5], experiential learning [29], and inquiry-based learning [13].

An inquiry-oriented approach to science education has two valuable effects. First, the literature on discovery-based learning points to improvements to student engagement and learning through participation with such a pedagogical approach [42, 46, 48]. Second, inquiry and discovery are valuable in and of themselves because they are authentic representations of participation in real scientific research. One goal of early science education is to stoke students' mastery of science and interest in science careers [43]; toward this end, participation in an authentic exercise is a valuable learning experience on its own [21].

However, these constructionist learning approaches are not without valid criticisms. Kirchner, Sweller, & Clark argue, from both theoretical and empirical viewpoints, that purely unguided or minimally-guided instructional approaches are insufficient [27]. Rather, they suggest that guided instruction is critical early in the learning process. Guided instruction, they argue, provides the learner with the foundational knowledge and skills necessary to begin to drive their own discovery and learning. Other studies have similarly corroborated the weakness of discovery-oriented approaches in certain settings and domains [28, 32].

The work presented here addresses this need for guided instruction in the context of scientific inquiry, modeling, and discovery. The objective of this research is to teach students the metacognitive process of inquiry-driven modeling through interaction with a set of intelligent agents embedded in a software environment. We have developed a series of exploratory learning environments that enable students to investigate complex ecological phenomena [23, 50] which leverage inquiry-driven teaching. Students demonstrated significant improvement in deep understanding after using these environments [17], but we also observed the kinds of weaknesses noted above. Learners often exhibited subpar investigative processes, and while their deep understanding of the systems improved, the actual process of inquiry and modeling did not.


*IUI 2015,* March 29–April 1, 2015, Atlanta, GA, USA.

http://dx.doi.org/10.1145/2678025.2701398.

Informed by these experiences, we have augmented our most recent exploratory learning environment, the Modeling & Inquiry Learning Application (MILA), with an extension to provide guided instruction directly in the context of a modeling and inquiry activity. This extension, MILA–Tutoring (MILA–T), supplies five distinct intelligent agents, each mimicking a particular functional role that a teacher traditionally plays in a classroom. In this paper, we first present the design of five pedagogical agents, the ways in which they track student behavior over time, the feedback that they provide, and the way in which they interact with teams during the inquiry-driven modeling process. We then present the design and results of an experiment with 238 middle school science students working in groups of two or three to investigate and construct a model of an ecological phenomenon. We present the differences in behavior seen based on interaction with the tutoring system and deeper analysis of the role that these tutoring agents played in the inquiry-driven modeling process.

## RELATED WORK

This research builds on prior research in a variety of communities, including intelligent tutoring systems, exploratory learning environments, artificial intelligent in education, intelligent user interfaces, and learning sciences. Most prominently, however, this work represents the union of two communities: modeling and inquiry in science education and metacognitive tutoring.

### Modeling and Inquiry in Education

Significant attention has been devoted to developing approaches to science education that encourage participation in authentic scientific practice. This kind of teaching strategy leverages the appeal of constructionist learning approaches described previously as well as exposes students to the authentic way in which science is performed in the real world. Such directions have been encouraged by both policy makers [33] and researchers [12, 30, 42]. Generally speaking, scientific models are representations of the system and are used to make predictions about the system for further analysis [34]. Models come in many different forms, including descriptive, prescriptive, conceptual, simulative, causal, dynamic, mechanistic, explanatory, visual, and more [8, 18]. For this research, we focus on models that are explanatory and mechanistic [17].

Scientific inquiry is tightly connected with scientific modeling. Inquiry is the process by which scientists or learners examine the domain or phenomenon that they are investigating, find useful observations or data, and use that data to test and expand their understanding of the system [43]. A scientific model, in this sense, serves as the destination for newly-unveiled data, as well as an organizing tool for determining future areas of inquiry [34]. Scientists use models to make predictions for what observations they would expect if the model is accurate; they then gather that data as part of the inquiry process and use it to corroborate the model's predictive power or elaborate on the model's mechanistic explanations [52].

Educational approaches that unite modeling and inquiry have been used extensively in science education [24, 43, 52]. However, these discovery-based approaches are subject to the same challenges noted previously [27, 28, 32]. Implementing guided instruction in these types of interventions presents a pragmatic difficulty: it is difficult for a single teacher to monitor and guide the open-ended inquiry and discovery of multiple students or groups of students at once. This difficulty is exacerbated in inquiry and modeling by the more general difficulties with teaching metacognitive skills, which have previously been identified by the metacognitive tutoring community.

### Metacognitive Tutoring

Metacognitive tutoring is an extension of cognitive or intelligent tutoring. Whereas intelligent tutoring systems typically address cognitive skills [e.g. 49, 53], metacognitive tutoring initiatives attempt to construct intelligent agents that teach students metacognitive skills like self-regulated learning [2], self-explanation [9] and help-seeking [1]. These kinds of metacognitive skills have been identified as one of the most crucial learning goals of early education [3, 6, 11]. However, teaching metacognition has a number of unique challenges. Roll et al. 2007 provides an overview of many of these unique challenges [40]. Metacognitive skills tend to be domain-independent, but must be taught within a specific domain. Students have a natural tendency to emphasize the domain-specific learning rather than the metacognitive skills. Metacognitive skills are also difficult to teach explicitly. Because metacognition occurs within the mind of the reasoner, there is no inherent externally observable behavior or skill from which to learn. In the following section, the nature of inquiry and modeling as metacognitive skills is presented; however, we may also see the connection between inquiry, modeling, and metacognition by the shared difficulties both present. Like other metacognitive skills, inquiry and modeling are difficult to teach because they exist largely in the mind of the learner or scientist and because there is a tendency to focus on the domain knowledge rather than the metacognitive skill.

## INQUIRY-DRIVEN MODELING

This research aims to teach students the process of inquiry-driven modeling within an authentic activity in ecological investigation. We define 'inquiry-driven modeling' as a particular type of the modeling and inquiry process in which the modeling process is driven by the results and data uncovered during inquiry activities; in turn, the resultant model helps structure and direct the continued inquiry activities. Other initiatives have taught inquiry [26, 48] and modeling [4, 47] separately, but this work teaches them together as directed by the literature on inquiry and modeling in authentic scientific research [34].

The case for inquiry-driven modeling as a metacognitive skill is derived from three connections. First, inquiry-driven modeling meets the definition of a metacognitive skill. Metacognition, as defined by Weinert 1987, is "cognition about cognition; that is, it refers to second-order cognitions: thoughts about thoughts, knowledge about knowledge or reflections about actions" [51]. The target of an inquiry-driven modeling task is one's own knowledge or understanding, aligning with the definition of a metacognitive skill. Second, inquiry-driven modeling is in many ways a local instantiation of the broader self-regulated learning process. Inquiry-driven modeling takes the general principles of self-regulated learning [2, 7, 11, 36], such as planning, self-monitoring, strategizing, and self-assessment, and deploys them in a particular domain with an additional set of rules, standards, and practices. Third, prior research has articulated the nature of inquiry and modeling as metacognitive skills. Most notably, White & Frederiksen explored this issue by initially using metacognition as a way of creating educational interventions grounded in inquiry and modeling [41, 52]. This work develops the idea of "metamodeling" knowledge, which is an understanding not simply of the process of modeling, but also to the role, function, and need for modeling in scientific inquiry. Metacognitive tutoring has also been applied previously to the inquiry phase of the process [16].

### Process of Inquiry-Driven Modeling

The objective of this research is to teach students the metacognitive process of inquiry-driven modeling. In order to do so, we must first articulate a desirable process of inquiry-driven modeling. Based on significant existing research on inquiry and modeling in both education [39, 43, 44, 52] and science [10, 34], and supported by our own experience with inquiry and modeling in our exploratory learning environments in the past [17, 23, 50], we have developed a model of the process of inquiry-driven modeling, as shown in Figure 1. In this process, the learner (whether a student learning about ecology in a classroom or a scientist learning about the world in an authentic research setting) starts off by observing and describing some phenomenon to investigate. They then propose one or more hypotheses that could explain this phenomenon (although some researchers suggest scientists gather information

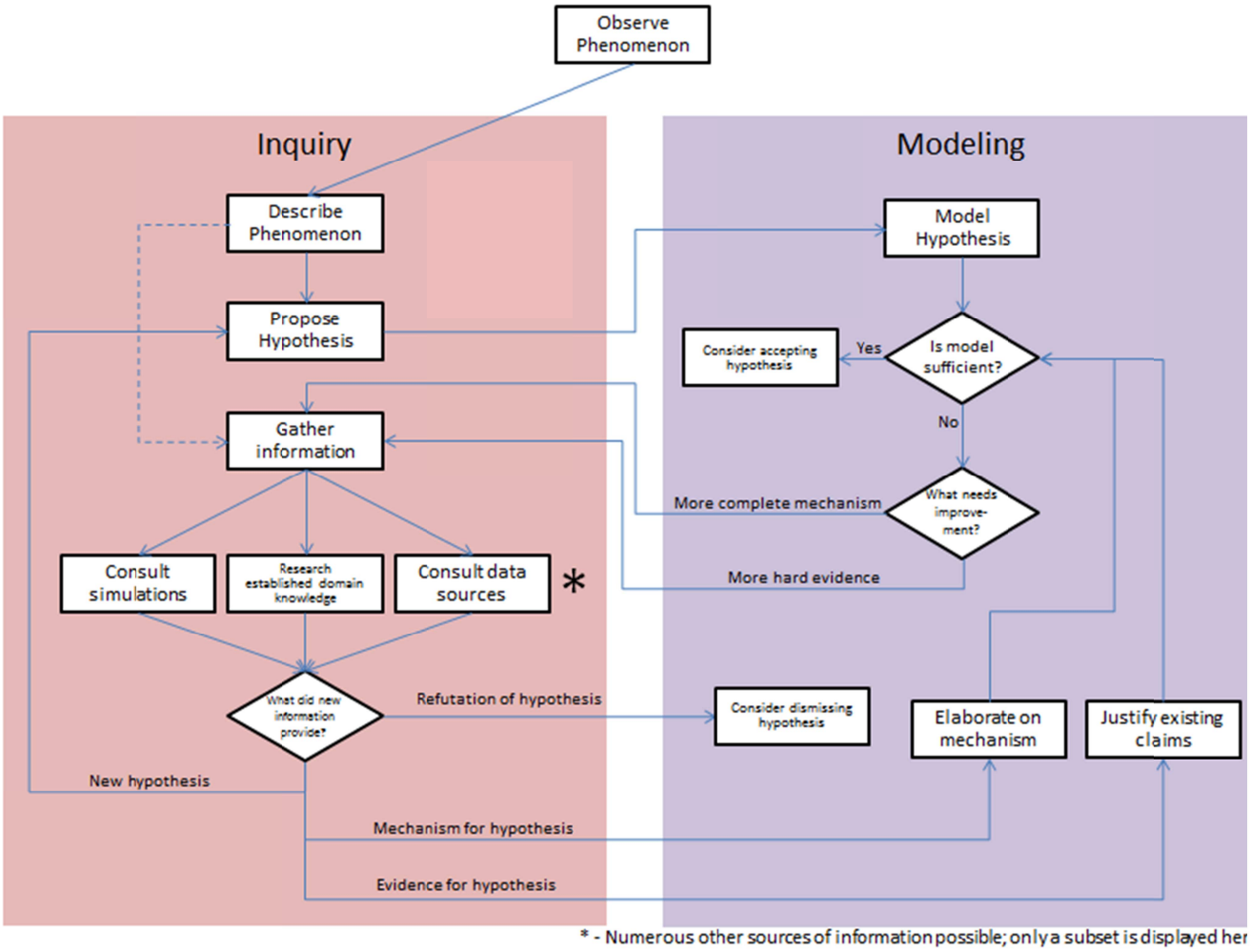


**Figure 1: A model of a desirable process of inquiry-driven modeling.**

before developing hypotheses [35], which is also supported by this model). These hypotheses then become preliminary models of the phenomenon. Using these models, the learner determines what is needed to confirm or expand the model. This leads to information-gathering in the world, which brings new information back to the model. Based on this new information, the model is altered, either dismissing it if the new information contradicts the model, expanding it if the new information provides mechanistic information, or strengthening it if the new information confirms predictions that the model had previously made. This process then continues to further elaborate and strengthen the model. It is important to note that the model presented in Figure 1 is not suggested as an ideal representation of the process, but rather just one faithful formalization of the literature on the inquiry and modeling process.

This model of a desirable inquiry-driven modeling process connects to the advice given in the metacognitive tutoring community that learning goals must be made explicit and explicitly communicated to the learners [40]. However, this process presents many of the learning challenges noted in the existing literature on modeling and inquiry in education and on metacognitive tutoring, as referenced in the Related Work section above. In a traditional classroom, all these difficulties are exacerbated even further by the need for a single teacher to provide guidance and instruction to multiple students or groups of students at once. This process is inherently explorational but demands guided instruction for productive learning [27], and a single teacher often will not be able to guide multiple groups exploring in different directions simultaneously. For these reasons, this research uses a metacognitive tutoring system to provide targeted, guided instruction to groups of students directly in the discovery context.

## MODELING AND INQUIRY LEARNING APPLICATION

The inquiry-driven modeling in this research takes place in an exploratory learning environment called the Modeling & Inquiry Learning Application (MILA). In order to understand the nature of the tutoring and inquiry-driven modeling that takes place in this environment, it is first necessary to understand the nature of the models that students construct within this environment.

Figure 2 shows the main MILA window. In the top left, teams write their description of the phenomenon that they are trying to describe. Teams then propose one or more hypotheses; each of these hypotheses then becomes a model of how that hypothesis could lead to the phenomenon. Teams may also use the left sidebar to launch simulations, take notes, and dismiss models they no longer wish to consider. Within the model, teams construct explanations comprised of nodes and edges. Each node features three parameters: the physical component of the system in the

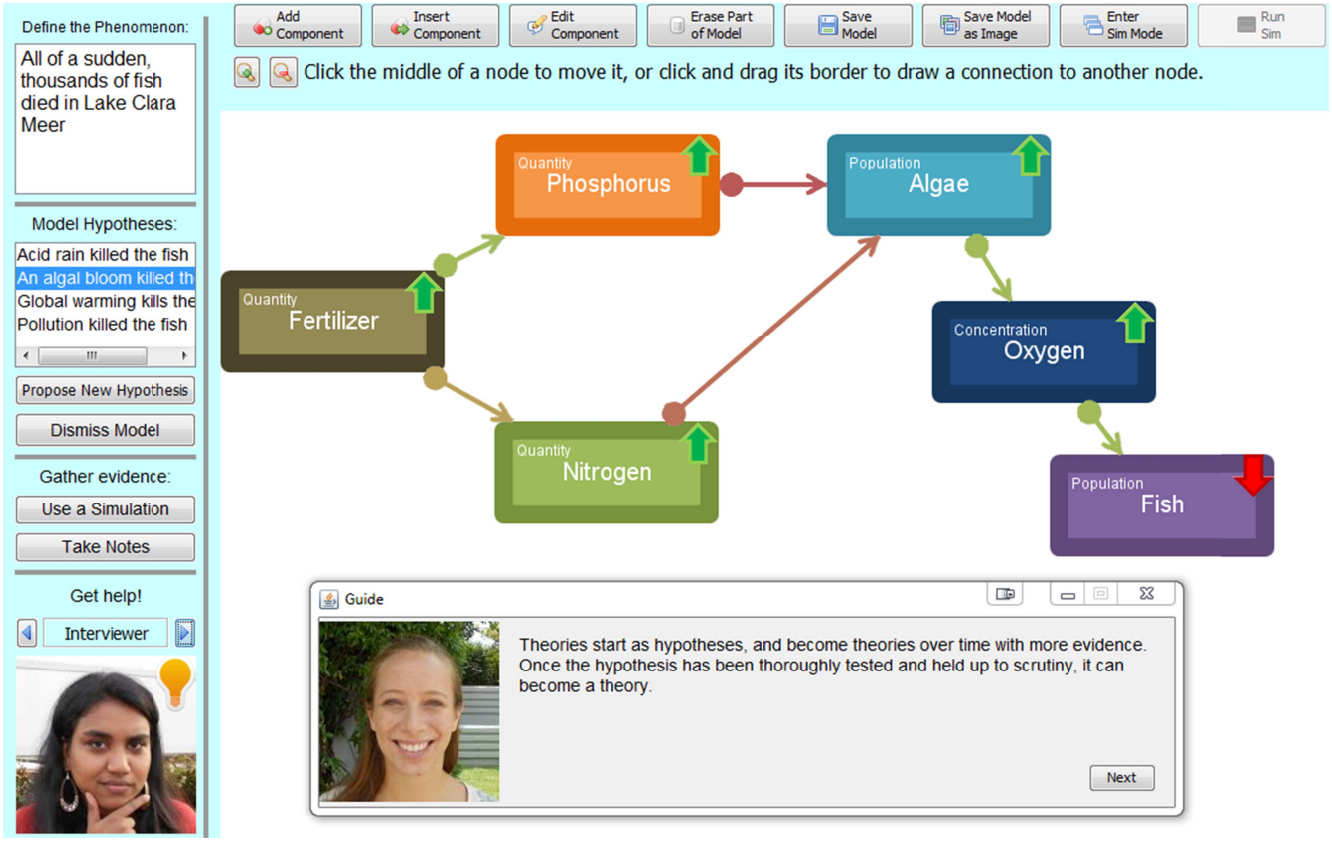


**Figure 2: The Modeling & Inquiry Learning Application being used to model a sudden, massive fish kill. MILA-T is shown in the bottom left, and pop-up feedback from the Guide is shown in the bottom middle.**

center (such as Fish or Phosphorus), the variable of the component in the top left (such as Population or Quantity), and the direction of change in the top right. Edges between nodes are causal; each trend causes the next one. For example, in this model, an increase in the quantity of fertilizer in a system causes an increased in the quantity of phosphorus and nitrogen. Students provide evidence for their models on the edges between nodes. On these edges, teams can write their explanation for why they believe a given connection is true and code it according to one of several categories of evidence, such as Logical Explanations and Similar System Observations. Based on the evidence provided, the color of the edge changes from red to orange to green; red signifies no evidence while green signifies ample evidence.

### MILA–TUTORING

In this research, MILA is augmented with a tutoring extension called MILA–Tutoring (MILA–T). MILA–T is comprised of five distinct pedagogical agents that monitor and respond to students' behavior within the software. The goal of these agents is to help the process of inquiry-driven modeling in which students participate converge more closely to the process presented in Figure 1. To do so, these agents specifically monitor for successful demonstration of the process, as well as for the errors described previously.

During engagement with MILA, the tutoring system is available in the bottom left corner of the window, as shown in Figure 2. Four tutors are visible: a Guide, a Critic, a Mentor, and an Interviewer. Each tutor plays a different functional role with regard to interaction with the team of students, but all are structured to teach the process of inquiry-driven modeling. A fifth tutor, the Observer, is invisible to students but operates in the background to provide information to the other tutors. Two of the tutors, the Guide and the Critic, will not provide feedback until it is solicited by the team, while the Mentor, Interviewer, and Observer are constantly monitoring the team's behavior and interrupting accordingly. When the Mentor and Interviewer wish to provide feedback to the team, they illuminate a light bulb icon on their avatar in the corner of the screen, as seen in Figure 2. Each agent follows a unique decision routine to intelligently select feedback to provide. The Guide and Critic run their routines whenever they are prompted for feedback by the team. The Mentor, Interviewer, and Observer run their decision routines after every action that the team takes, with a threshold constraining how often they will provide feedback to the team.

#### The Observer

The Observer constantly monitors the activity of the team and constructs three different assessments of the team's ability: a modeling assessment, an inquiry assessment, and an ecology assessment. Each model reflects a different aspect of the desired inquiry-driven modeling process. To construct the modeling assessment, the Observer watches the pattern of teams' construction and revision activities over time, checking for a willingness to rescind earlier conjectures, dismiss earlier hypotheses, and propose new explanations. To construct the inquiry assessment, the Observer monitors the quality and quantity of evidence teams provide in support of their explanation, including the degree to which articulating evidence is embedded in the model construction process. To construct the ecology assessment, the Observer checks for the presence of certain desired components and relationships, such as the team's ability to include invisible components in their model like chemicals and microscopic organisms. Each assessment is constructed of a number of lower-level criteria, such as a measure of the tendency of a team to rely on weak forms of evidence or the willingness of a team to remove or revise past portions of their explanation. These criteria are then used to establish an overall level of ability with the target skill for the other tutors to check (although the other tutors may also check the individual criteria). One example of a rule that the Observer uses to construct its assessment of the team's ability is:

> **IF:** Students have 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 assessment.

These models are provided to the other tutors to help them establish the team's ability and provide the proper feedback. The Observer also has a rudimentary model describing the team's interaction with the various tutors themselves so that the other tutors may provide feedback praising teams' willingness to seek help or encouraging them to use the tutoring system.

#### The Guide

The Guide's functional role is to anticipate the questions that the team may want to ask and provide answers to those questions on demand. To accomplish this, the Guide constructs a list of questions that the team may want to ask, with answers (or a list of follow-up questions) prepared for each question. To do so, she iterates through a list of rules each containing a set of percepts. These percepts examine several sources: the Observer's current model of the team's ability across all three dimensions, the current classroom context, and the current status of the team's models. Each rule maps a particular set of percepts to a question or set of questions the Guide may offer to the team; offering the question to the team is thus the action that the rule selects in response to the particular set of percepts. One example of a rule that the Guide uses to choose a question is:

> **IF:** Students have begun to add evidential justifications to their models; OR: (Students have reached a point in the curriculum when evidence has been introduced; AND: Students have not yet reached efficacy with evidential justifications according to the

Observer's model); OR: The students' 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?".

As she checks her rules, the Guide compiles a list of questions to offer. Once this compilation is done, the Guide offers the questions to the team, and they may select a question to which to receive an answer. Figure 2 shows the Guide providing an answer to a team's question about the difference between a hypothesis and a theory.

The Guide is equipped with dozens of questions to offer to teams. These questions range from novice-oriented questions such as simple information on interacting with the software to expert-oriented questions on evaluating explanations and establishing scientific theories. The Guide is also equipped with targeted questions that identify and address specific actions or features of the team's model; for example, if the team constructs a model reliant on logical explanations, the Guide offers a question about how one might gather data to test such explanations.

### The Critic

The Critic's functional role is to provide teams with feedback on the current quality of their model. Teams are encouraged to consult with the Critic when they are unsure of how to proceed or believe their explanation is sufficient. Like the Guide, the Critic is equipped with a set of rules that determine what feedback he selects and provides. The rules are ordered from basic feedback to advanced feedback. When called upon, the Critic iterates over each rule and checks the percepts associated with the rule; if the percepts trigger a match to the current state of the team's models or the Observer's assessment of the team's ability, the feedback associated with the rule is added to a pool of potential feedback to provide. Once an adequate selection of feedback is attained (~5 different suggestions), the Critic randomly selects one and displays it to the team. In this way, if the team asks for feedback multiple times, they receive different suggestions and can proceed even if the Critic provides a piece of feedback they cannot currently address. One example of a rule the Critic uses to select a piece of feedback is:

**IF:** Students have not yet written a phenomenon definition; AND: Students have 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!"

Like the Guide, the Critic's feedback is based partially on the Observer's model of the team's ability and partially on the current state of the model that the team is currently constructing. For example, if a team has previously demonstrated an understanding of desirable forms of evidence (as seen in the Observer's assessment of the team's ability), the Critic will skip to more advanced feedback even if the current model retains some more basic weaknesses. In this instance, the Critic infers that the team is aware of this weakness because they have demonstrated an understanding of it in the past (although the Critic will still default to this feedback if it cannot find any more advanced feedback).

### The Mentor

The Mentor similarly monitors for weaknesses in the team's modeling and evidence, but unlike the Critic, the Mentor will intervene and provide the team with unsolicited feedback. Thus, the Mentor plays the functional role of providing teams with feedback even when they are not soliciting help. The Mentor is comprised of a list of rules that he checks every time the team performs an action in the software. If the Mentor identifies a piece of feedback that he has not previously provided to the team, and if he has waited a certain period of time since the previous piece of feedback was provided, he interrupts the team to provide the feedback. If no such piece of feedback is identified, the Mentor remains dormant. One example of a rule used by the Mentor to identify feedback for the team is:

**IF:** The student has recently dismissed one of their models; AND: The student had not yet demonstrated proficiency with proposing and dismissing models according to the Observer's model of the student; AND: The student 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."

The Mentor is primarily comprised of pairs of rules each targeting certain portions of the Observer's assessment of the team's ability. If the team demonstrates a certain desired skill, the Mentor will praise the team's ability; if the team has not demonstrated that skill within a certain period of time, the Mentor will describe the skill and its value to the team.

### The Interviewer

The Interviewer's function is to ask the team questions that they ought to learn to ask themselves. Every time the team performs an action, she checks if the action and other present information match the percepts for one of her rules. If so, she provides the question to the team, along with a

text box to use to answer the question. An example of a rule used by the Interviewer to select a question is:

> **IF:** Students have 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?", followed by, "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!"

The Interviewer checks every action that the team performs and intervenes at critical junctures to ask the team to explain its reasoning. In this way, the Interviewer aims to encourage reflective learning by explicitly asking the team to reflect on their reasoning at critical junctures of the inquiry-driven modeling process. When prompted to reflect by the Interviewer, she provides the team with a text box in which to write their answer; this information is stored for subsequent analysis.

### Interactions Among Tutors

The tutors are also symbiotic in that they also reference one another during their interactions with the teams. The Critic, for example, suggests the team consult the Guide if they are unclear about why he is emphasizing strong pieces of evidence or more elaborate mechanisms. The Guide encourages teams to pay attention to the feedback from the Mentor in order to discern whether they are successfully executing the inquiry-driven modeling process. The Mentor promotes heavy use of the Critic when teams appear to stagnate in development of their explanations as a way of discerning areas for improvement. Both the Guide and the Mentor also help frame the Interviewer's questions as reflective exercises rather than summative assessments. This symbiotic relationship may also explain the observation that interaction with the tutoring system most often spawns further interaction with the tutoring system. As noted below, 56.34% of all interactions with one of the tutors were followed by another interaction with the tutoring system.

## EXPERIMENTAL DESIGN

MILA and MILA–T were deployed in two-week unit in 7th grade life science classrooms. During this unit, students were broken into teams of two or three and completed two projects. The first project (the "Learning" project) was given four 50-minute periods; in this project, students were asked to explain a sudden, massive fish kill in Lake Clara Meer. The second project (the "Transfer" project) was given one 50-minute period; in this project, students were asked to explain Atlanta's record-high temperatures over the past 20 years. In addition to these five periods, students spent two periods gathering data and completing lab exercises without MILA and two periods completing assessments.

Two teachers participated in the intervention, each with five classes. For each teacher, two classes were assigned to a control group and three classes were assigned to an experimental group. In the control group, 34 teams of students (99 total students) used MILA without MILA–T. In the experimental group, 50 teams of students (138 total students) used MILA with MILA–T during the Learning project and MILA without MILA–T during the Transfer project (only 47 experimental teams completed the Transfer project). This design enables identification of how teams interact differently while receiving feedback from the tutoring system and how teams' behavior changes in future projects based on prior interaction with a tutoring system.

We collected several types of data during this intervention. For this analysis, the most significant data are the interaction logs taken from each team. At the conclusion of the intervention, these logs were gathered together and separated by group (control or experimental) and project (Learning or Transfer). These logs then became the primary data source for the analysis outlined below.

## ANALYSIS

We conducted two analyses on these interaction logs. First, we compared the interaction logs between the control and experimental students to see how interaction with MILA–T altered students' modeling and inquiry process, both during interaction with MILA–T (the Learning project) and after MILA–T was disabled (the Transfer project). Then, we analyzed the way in which the tutoring system was used by the experimental group during the Learning project. In order to perform both these analyses, the raw interaction logs were processed into Markov Chains mapping the software interactions into phases of the inquiry-driven modeling process described in Figure 1.

### Markov Chains

To analyze the log data, we developed Markov chains describing the patterns of interaction in which teams engaged. Markov chains are mathematical systems that summarize transitions amongst a number of distinct states in a state space [25]. Markov chains are characterized as memoryless; it is inferred that next state chosen in the state space is determined by a probability function taking only the current state as an argument [31]. Although we would infer that a longer history of interactions likely helps determine the next state in a sequence in our analysis, for the purpose of this analysis Markov chains provide a useful device for examining differences based on the presence of the metacognitive tutoring system. Markov chains differ from hidden Markov models in that Markov chains allow the states themselves to be identified separately from the data; hidden Markov models derive the states from patterns in the data [15]. Markov chains were chosen rather than hidden Markov models to allow for more direct mapping to the inquiry-driven modeling process described earlier.

Part of one of these Markov chains is shown in Figure 3. Given the complexity of these Markov chains (twelve states with almost 100 notable edges), the chain is presented here as a table rather than a more traditional set of nodes and

| | Number of Instances | Constructing Evidence | Constructing Model | Consulting Tutor | Dismissing Model | Proposing Hypothesis | Reconsidering Model | Revising Evidence | Revising Model | Taking Notes | Using Simulation | Writin |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 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% |

**Figure 3: One of the Markov models used to analyze teams' patterns of interaction with MILA and MILA-T. Due to the number of activities, the chain is shown as a table.**

edges. Along the left and along the top, the twelve activities in which students engage during the inquiry-driven modeling process are shown; these activities summarize lower-level software interactions and map those interactions to the inquiry-driven modeling process presented in Figure 1. Differentiating construction and revision of evidence is done to capture modifications to prior ideas instead of merely ongoing additions to the model. Consulting Tutor is not a part of the inquiry-driven modeling process, but rather is a general activity to be performed whenever students are stuck, believe they are finished, or otherwise receive input. The rows represent the "previous" activity a team completed and the columns represent the "next" activity a team completed (except the "Number of Instances" column, which shows the number of observed instances of the activity for that row). In this way, each cell of the table represents the number of times teams followed the activity on the left with the activity on top. For example, at the intersection of the Constructing Model row and Constructing Evidence column is the number 5.07%; this means that 5.07% of the time that teams performed an action in the Constructing Model activity, they followed it with a Constructing Evidence activity. Similarly, the intersection between the Consulting Tutor row and the Consulting Tutor column shows the number 56.34%: this means that 56.35% of interactions with the tutoring system were followed by another interaction with the tutoring system.

These Markov chains are useful tools for analysis because they summarize the overall pattern of interaction performed by each group. Aggregating together the interaction logs from all teams in the control and experimental groups gives us expansive Markov chains summarizing thousands of activities, allowing us to compare the patterns of interaction between teams interacting with the tutoring system and those without it. However, merely comparing the patterns of interaction is of limited usefulness; the goal of interaction with the intelligent tutoring agents is to improve the team's inquiry-driven modeling. Toward that end, it is necessary to articulate what "improved" inquiry-driven modeling would look like in these models.

### Desired Improvements

The objective of this research is to teach students the metacognitive process of inquiry-driven modeling. As described previously, MILA–T gives students feedback on the modeling process, and the desire is to see students respond to this feedback by improving their inquiry-driven modeling process. In order to identify improvement, however, it is first necessary to operationalize what improvement would look like in terms of these Markov models. Based on the research and our prior experience in developing and deploying exploratory learning environments, we derived nine behaviors that would be indicative of improved inquiry-driven modeling that are captured by these Markov chains. For the purposes of describing the operationalization of these desired improvements, 'Using Simulation' and 'Taking Notes' are combined under the broader activity of 'Data Gathering Activities'. The nine differences we monitor for in the Markov chains are:

- Greater incidence of articulating the problem prior to modeling and inquiry. This is operationalized by a greater prevalence of a transition between Start and Writing Problem Definition (found at the intersection of the Start row and Writing Problem Definition column in the Markov chain in Figure 3).
- Greater incidence of using the results of data-gathering activities to construct new evidential justifications. This is operationalized as a greater prevalence of a transition between Data-Gathering Activities and Constructing Evidence.

- Greater incidence of using the results of data-gathering activities to revise previous evidential justifications. This is operationalized as a greater prevalence of a transition between Data-Gathering Activities and Revising Evidence.
- Greater incidence of using the results of data-gathering activities to refute previous hypotheses. This is operationalized as a greater prevalence of a transition between Data-Gathering Activities and Dismissing Model.
- Greater incidence of using the results of data-gathering activities to propose new hypotheses. This is operationalized as a greater prevalence of a transition between Data-Gathering Activities and Proposing Hypothesis.
- Greater incidence of revising the problem definition throughout the process. This is operationalized as a greater overall prevalence of the Writing Problem Definition activity (the percentage of all actions that fall into the Writing Problem Definition activity, not shown in the model in Figure 3).
- Greater incidence model revision. This is operationalized as a greater overall prevalence of the Revising Model activity (the percentage of all actions that fall into the Revising Model activity, not shown in the model in Figure 3).
- Greater incidence of model construction activities spawning further information-gathering activities. This is operationalized as a greater prevalence of a transition between Model Construction and Data-Gathering Activities.
- Greater incidence of note-taking. This is operationalized as a greater overall prevalence of Taking Notes from data-gathering outside MILA (the percentage of all actions that fall into the Taking Notes activity, not shown in the model in Figure 3).

These nine criteria are regarded as indicators that one team's metacognition is superior to another's. Even in the absence of acceptance of these criteria as indicative of improved metacognition, however, these nine criteria represent desirable improvements in the inquiry-driven modeling process.

**Comparison of Markov Chains**

This research hypothesizes that interaction with the metacognitive tutoring system will improve teams' inquiry-driven modeling process compared to teams that do not interact with the metacognitive tutoring system. For this analysis, improvement is operationalized as greater incidence of the nine desirable behaviors outlined previously. Thus, the goal of this analysis is to identify whether teams in the experimental group exhibited greater incidence of any of these nine desired behaviors compared to teams in the control group. To determine whether the experimental group was superior in any of these nine comparisons, we performed a two-tailed Z-test on each pair of numbers derived from the Markov chains. For example, in order to test whether the experimental group demonstrated a great incidence writing problem definitions prior to performing any modeling and inquiry activities during the Learning project, we performed a two-tailed Z-test comparing the value at the intersection of the Start row and Writing Problem Definition column in the experimental group's Markov chain (in this case, 90.00%) with the corresponding value in the control group's Markov chain (in this case, 88.24%).

This repeated Z-test approach raises the odds of a Type I (false positive) error. To account for this, we used $\alpha = 0.01$ as the threshold for accepting the results of any one of the Z-tests. We then performed a Bernoulli (or binomial) trial to determine the number of expected false positives with nine trials and $p = 0.01$. This Bernoulli trial revealed a 91.4% chance of no false positives, an 8.3% chance of one false positive, and a 0.3% chance of more than one false positive. 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).

During the Learning project, three of the nine criteria showed statistically significant differences between the control and experimental groups. Teams in the experimental group demonstrated an increased propensity to revise their problem definitions over time ($p < 0.001$, $Z = 4.102$) and to revise their models over time more generally ($p < 0.01$, $Z = 2.647$). The control group demonstrated an increased propensity to take notes ($p < 0.01$, $Z = -2.89$). Based on the aforementioned Bernoulli trial, we can conclude that no more than one of these differences was a false positive, and therefore there did exist statistically significant improvements by the experimental group in their propensity to revise either problem definitions or models over time. Given that one of the three differences presented improved performance by the control group, we must stop short of claiming that the experimental group was more generally superior. Nonetheless, we may conclude that the experimental group was superior in at least one dimension.

During the Transfer project, three of the nine criteria showed statistically significant differences between the control and experimental groups. Teams in the experimental group demonstrated an increased propensity to revise their models over time ($p < 0.01$, $Z = 2.716$) and to take notes ($p < 0.0001$, $Z = 3.593$). Teams in the control group demonstrated a significantly increased propensity to revise their problem definitions over time ($p < 0.01$, $Z = -3.511$). Interestingly, the control and experimental groups alternated areas of superiority between the two projects. Considering the Bernoulli trial described, we can conclude that a minimum of two of these observed differences were

not false positives. The experimental group was superior to the control group in either their propensity to revise their models over time or their propensity to take notes. Although we again must stop short of claiming that the experimental group was more generally superior, we can conclude that the experimental group was superior in at least one dimension.

The results of the Bernoulli trials show that given three successes, two must not be false positives; thus, for each project, either the experimental group was superior in two dimensions or the control and experimental groups were each superior in one dimension. Given that either possibility features the superiority of the experimental group in at least one dimension, we may conclude that the experimental group was superior in at least one dimension during both the Learning and Transfer projects.

### Examination of Tutor Usage Patterns

Following the previous analysis of the difference in inquiry-driven modeling processes between control and experimental groups, we conducted a follow-up analysis more narrowly on the role that MILA–T played in the experimental group during the Learning project. The goal of this analysis is to determine the nature of teams' interaction with the tutoring system and to construct a case for how the tutoring system impacts the teams' inquiry-driven modeling process, with the hypothesis that interaction improves their execution of the process. Using an additional Markov chain of the transitions amongst the activities including each tutor individually, we identified a number of patterns to the usage of MILA–T among experimental teams. These observations are qualitative given the lack of an alternate pattern of engagement with the tutors against which to test these values.

The first observation is that interaction with MILA–T was deeply embedded in experimental teams' interactions with MILA during the Learning project. 24.64% of all interactions during the Learning project fell into the Consulting Tutor activity, with 1769 total individual instances of teams consulting the tutoring system. Each of these instances represent the team receiving feedback from a tutor or giving an answer to one of the Interviewer's questions. The second observation is that tutor interaction most often begets further tutor interaction. 56.34% of tutor interactions are followed by an additional tutor interaction, suggesting that teams of students seek multiple pieces of feedback. This most often was seen in interactions with the Critic as students requested different feedback.

The third significant observation is that after exiting a loop of tutor interactions, teams most often transition to either model construction or evidence construction. This mirrors the feedback teams most often receive from the tutoring system. Much of the tutoring system's targeted feedback attempts to support teams in transitioning toward mature modeling and inquiry behaviors characterized by varied and strong evidence in support of their models and thorough mechanisms explaining the chain of events leading from their hypotheses to the phenomenon. In both these cases, the solution is to construct additional portions of the model by either providing more evidence or elaborating on the mechanism. The tutors most often ground this feedback in the teams' current explanations, pointing out when they are relying too heavily on simple logical explanations and observations. In these instances, though, removing prior evidence is not encouraged; instead, teams are encouraged to corroborate their logical explanations with stronger evidence. These observations are corroborated by other Markov chains at the level of the specific tutor. The Mentor's feedback is most often followed by model construction (15.45%) and model revision (10.30%), while the Critic's feedback is most often addressed by evidence construction (15.63%) and evidence revision (2.95%).

## CONCLUSIONS

This paper presents a set of intelligent tutoring agents that teach students an authentic metacognitive process of inquiry-driven modeling. Based on the results seen here, we see evidence that engagement with a metacognitive tutoring system during an inquiry-driven modeling activity improved students' participation in the process in certain specific ways. Although the Bernoulli trial leaves open the possibility that certain improvements may be false positives, the repeated improvement in propensity to revise models over time in the Learning and Transfer projects suggests that teams both acted on and internalized the feedback received from the tutoring system. The qualitative analysis of interaction with the tutoring system, which showed tutor interactions were most often followed by construction and revision activities, further supports that interaction with the tutors encouraged further construction and revision. This suggests more frequent iteration through the inquiry-driven modeling process, which may be interpreted as decreased propensity to accept an explanation of the phenomenon too early. Further analysis is required to validate these ideas; the data presented here merely shows that teams receiving feedback from the tutoring system show an increased propensity to revise their models over time rather than adhere to their initial conjectures.

This analysis provides one perspective on the results of this intervention. We have also found that engagement with MILA–T improves students' dispositional orientation toward science and science careers [21], and improves the final explanations of the target phenomena that teams generate [20]. We are also exploring the impact of other features of MILA, such as MILA–S, a system for generating simulations from the conceptual models developed within MILA [22].

## ACKNOWLEDGMENTS

We are grateful to Lara Catall, Vickie Bates, and Kristina Strickland for teaching in this intervention and to Nicolas Papin and Rochelle Lobo for their support in developing MILA, MILA–T, and MILA–S.

## REFERENCES


1. Aleven, V., McLaren, B., Roll, I., & Koedinger, K. (2006). Toward metacognitive tutoring: a model of help seeking with a cognitive tutor. *International Journal of Artificial Intelligence and Education 16*(2), 101-128.
2. 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.
3. Baker, L., & Brown, A. L. (1984). Metacognitive skills and reading. *Handbook of reading research, 1*, 353-394.
4. 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.
5. Blumenfeld, P. C., Soloway, E., Marx, R. W., Krajcik, J. S., Guzdial, M., & Palincsar, A. (1991). Motivating project-based learning: Sustaining the doing, supporting the learning. *Educational psychologist, 26*(3-4), 369-398.
6. Bransford, J., Brown, A., & Cocking, R. (2000). *How people learn: Brain, mind, experience, and school*. Washington, D.C.: National Academy Press.
7. Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of educational research*, *65*(3), 245-281.
8. Clement, J. (2008). *Creative Model Construction in Scientists and Students: The Role of Imagery, Analogy, and Mental Simulation*. Dordrecht: Springer.
9. Conati, C., & Vanlehn, K. (2000). Toward computer-based support of meta-cognitive skills: A computational framework to coach self-explanation. *International Journal of Artificial Intelligence in Education, 11*, 389-415.
10. Crawford, S., & Stucki, L. (1990). Peer review and the changing research record. *Journal of the American Society for Information Science, 41*(3), 223-228.
11. Dweck, C. (2000). *Self-theories: Their role in motivation, personality, and development*. Philadelphia, PA: Psychology Press.
12. 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.
13. Edelson, D. C., Gordin, D. N., & Pea, R. D. (1999). Addressing the challenges of inquiry-based learning through technology and curriculum design. *Journal of the Learning Sciences, 8*(3-4), 391-450.
14. Gallagher, S. A. (1997). Problem-based learning. Journal for the Education of the Gifted, 20(4), 332-62.
15. Ghahramani, Z. (2001). An introduction to hidden Markov models and Bayesian networks. *International Journal of Pattern Recognition and Artificial Intelligence*, *15*(01), 9-42.
16. Gobert, J., Sao Pedro, M., Toto, E., Montalvo, O., & Baker, R. (2011). Science ASSISTments: Assessing and scaffolding students' inquiry skills in real time. Paper presented at The Annual Meeting of the American Educational Research Association, April, 2011, New Orleans, LA.
17. 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 Complex Systems: A Reflection on 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*.
18. Griffith, T., Nersessian, N., & Goel, A. (2000). Function-follows-Form: Generative Modeling in Scientific Reasoning. In *Proceeds of the 22nd Cognitive Science Conference*.
19. Jacobs, G. (1992). Hypermedia and discovery-based learning: a historical perspective. British Journal of Educational Technology, 23(2), 113-121.
20. Joyner, D. (2015). *Metacognitive Tutoring for Inquiry-Driven Modeling* (Doctoral dissertation). Georgia Institute of Technology, Atlanta, GA.
21. 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*. Honolulu, Hawaii.
22. 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*. Haifa, Israel. 289-298.
23. Joyner, D. A., Goel, A., Rugaber, S., Hmelo-Silver, C., & Jordan, R. (2012). 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*, Athens, GA.
24. Kaberman, Z., & Dori, Y. J. (2009). Question posing, inquiry, and modeling skills of chemistry students in the case-based computerized laboratory environment. *International Journal of Science and Mathematics Education, 7*(3), 597-625.
25. Kemeny, J. G., & Snell, J. L. (1960). *Finite markov chains* (Vol. 356). Princeton, NJ: van Nostrand.
26. 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.
27. 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.

28. 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.
29. Kolb, D. A. (1984). *Experiential learning: Experience as the source of learning and development.* Englewood Cliffs, NJ: Prentice-Hall.
30. Lajoie, S., Lavigne, N., Guerrera, C., & Munsie, S. (2001). Constructing knowledge in the context of Bio World. *Instructional Science, 29*, 155-186.
31. 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* (pp. 57-70). Springer Berlin Heidelberg.
32. Mayer, R. E. (2004). Should there be a three-strikes rule against pure discovery learning?. *American Psychologist*, *59*(1), 14.
33. National Research Council. (1996). *National science education standards.* Washington, DC: National Academy Press.
34. 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.
35. Nersessian, N. (2008). *Creating Scientific Concepts*. Cambridge, MA: MIT Press.
36. Palinscar, A. S., & Brown, A. L. (1984). Reciprocal teaching of comprehension-fostering and comprehension-monitoring activities. *Cognition and instruction*, *1*(2), 117-175.
37. Papert, S. (1980). Mindstorms: Children, computers, and powerful ideas. New York: Basic Books.
38. Piaget, J. (1950). The Psychology of Intelligence. New York: Routledge.
39. Razzouk, R. & Shute, V. J. (2012). What is design thinking and why is it important? *Review of Educational Research, 82*(3), 330-348.
40. 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).
41. 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 Verlag.
42. Schmidt, H. G. (1998). Problem-based learning: Does it prepare medical students to become better doctors? *The Medical Journal of Australia 168.* 429-430.
43. Schwarz, C. V., & White, B. Y. (2005). Metamodeling knowledge: Developing students' understanding of scientific modeling. *Cognition and Instruction*, *23*(2), 165-205.
44. Schwarz, C. V., Reiser, B. J., Davis, E. A., Kenyon, L., Achér, A., Fortus, D., ... & 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.
45. 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.
46. 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.
47. Ting, C. Y., Zadeh, M. R. B., & Chong, Y. K. (2006, January). A decision-theoretic approach to scientific inquiry exploratory learning environment. In *Intelligent Tutoring Systems* (pp. 85-94). Springer Berlin Heidelberg.
48. 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 Behaviors 21.* 671-688.
49. 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.
50. Vattam, S., Goel, A. K., Rugaber, S., Hmelo-Silver, C., Jordan, R., Gray, S., & Sinha, S. (2011). Understanding Complex Natural Systems by Articulating Structure-Behavior-Function Models. *Educational Technology & Society, Special Issue on Creative Design, 14*(1), 66-81.
51. Weinert, F. E. (1987). Introduction and overview: Metacognition and motivation as determinants of effective learning and understanding. *Metacognition, motivation, and understanding.* 1-16.
52. White, B. & Frederiksen, J. (1998). Inquiry, modeling, and metacognition: Making science accessible to all students. *Cognition and Instruction, 16*(1). 3–118.
53. Woolf, B. P. (2010). *Building intelligent interactive tutors: Student-centered strategies for revolutionizing e-learning.* Morgan Kaufmann.