Joyner, D. & Goel, A. (2015). Organizing Metacognitive Tutoring Around Functional Roles of Teachers. In *Proceedings of the 27th Annual Cognitive Science Society Meeting*. Pasadena, California.

# Organizing Metacognitive Tutoring Around Functional Roles of Teachers

**David A. Joyner (david.joyner@gatech.edu)**
**Ashok K. Goel (goel@cc.gatech.edu)**
Design & Intelligence Laboratory, Georgia Institute of Technology, Atlanta, GA 30338

**Abstract**

Metacognitive skills are critical in learning but difficult to teach. Thus the question becomes how can we facilitate metacognitive tutoring? We present an exploratory learning environment called MILA-T with embedded metacognitive tutors imitating five functional roles of teachers in classrooms. We tested MILA–T in a controlled experiment with 237 middle school students. We examine the impact of MILA–T on the models of a natural phenomenon constructed by the students. We find that students with access to MILA–T wrote better evidential justifications for their models, and thus, deliver better-justified models for the phenomenon. We also find that these improvements persisted during a transfer task. These results lend support for organizing metacognitive tutoring around the functional roles of teachers for supporting inquiry-driven modeling.

**Keywords:** Functional roles of teachers; intelligent tutoring systems, metacognition, metacognitive tutoring; scientific inquiry; scientific modeling.

## Introduction

Both learning scientists and emerging educational standards assert the need to teach authentic science to students from an early age (e.g. National Research Council 1996; Edelson 1997). Research in cognitive science describes scientific discovery as an iterative process entailing four related but distinct phases (Darden 1998; Nersessian 2008): model construction, use, evaluation, and revision. Thus, a model is first constructed to explain some observations of a phenomenon. The model is then used to make predictions about other aspects of the phenomenon. The model's predictions next are evaluated against actual observations of the system. Finally, the model is revised based on the evaluations to correct errors and improve the model's explanatory and predictive efficacy. Research in cognitive science also relates this process of scientific inquiry and modeling to metacognition (e.g. Clement 2008; Nersessian 2008; Schwarz et al. 2009; White & Frederiksen 1998): scientific inquiry and modeling is the metacognitive ability to reason over one's own understanding of a scientific phenomenon, construct an evidence-backed model of one's understanding, and use that model to inform further investigation into the system. This suggests the use of metacognitive tutoring to scaffold learning about inquiry-driven modeling.

However, metacognitive skills are generally difficult to teach (Roll et al. 2007), and teaching inquiry-driven modeling is no different. In past, exploratory learning environments (e.g. van Joolingen et al. 2005) and intelligent tutoring systems (e.g. Azevedo et al. 2009, 2010) have been successful in enhancing students' metacognition at least to a limited degree, indicating that it is in principle feasible to help develop metacognitive skills. However, it is not yet clear how to facilitate metacognitive tutoring, especially in exploratory learning environments for open-ended tasks such as inquiry-driven modeling: we still need to identify organizing principles for metacognitive tutoring.

One way to organize metacognitive tutoring is along the functional roles of teachers in science classrooms. Thus we developed a categorization of some of the functional roles of teachers: guide, critic, mentor, interviewer, and observer. Next, we developed an exploratory learning environment for inquiry-driven modeling (the Modeling & Inquiry Learning Application, or MILA), and a metacognitive tutoring system (MILA-T) consisting of metacognitive tutors imitating the roles of a guide, critic, mentor, interviewer, observer. Then, to evaluate if MILA-T scaffolded inquiry-driven modeling, we introduced the tutoring system to 237 middle school science students engaged in modeling a natural ecological phenomenon. We found that engagement with MILA–T led teams to developer better-justified models of scientific phenomena than teams without a metacognitive tutoring system. In addition we tested the efficacy of above learning on a transfer task. We also found that the improvements in model quality persisted even after the tutoring system has been disabled, showing that access to MILA–T actually improves students' inquiry-driven modeling and that those improvements transfer to a new task.

## MILA: Modeling & Inquiry Learning Application

MILA is an exploratory learning environment in which students working in small teams can participate in an authentic process of scientific modeling and inquiry. Teams investigate a natural phenomenon, posing multiple hypotheses that could explain the phenomenon, constructing models to provide mechanisms to those explanations, and supplying evidence to support those hypotheses and mechanisms. The MILA window is illustrated in Figure 1.

Models in MILA are made of nodes and edges. Each node in a MILA model represents a changing trend: for example, the first node in the chain shown in Figure 1 is that the Quantity of Fertilizer is Increasing, a trend in the system. These nodes are then linked together in a causal chain that explains some phenomenon: in Figure 1, the increasing fertilizer quantity leads to an increase in the quantity of phosphorus in the system. This leads to an increase in the population of algae, which leads to an increased concentration of oxygen, which leads to a decreased population of fish. Note that this model is incomplete: further mechanism is needed to show how the increased quantity of oxygen leads to a decreased fish population.

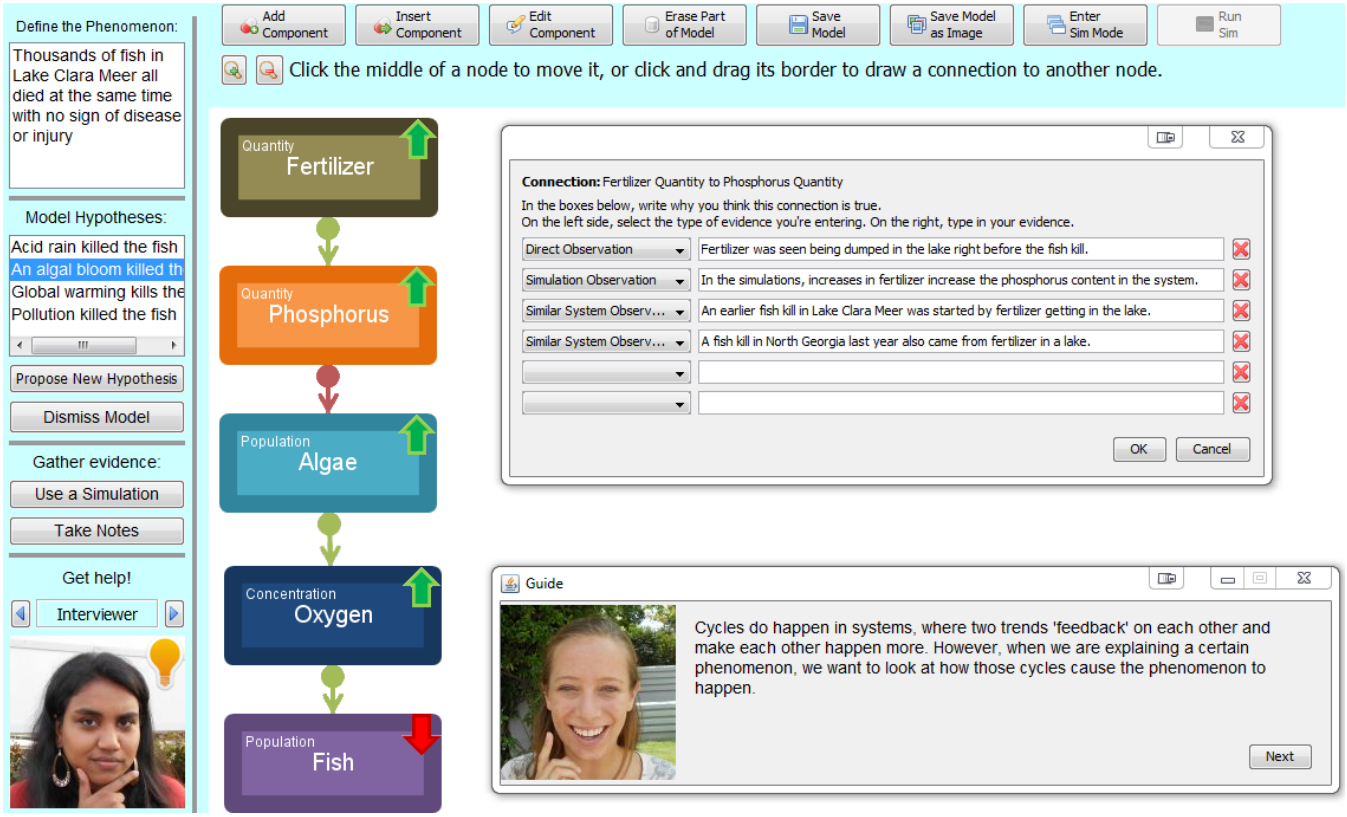


Figure 1: MILA and MILA–T. In MILA, teams describe a phenomenon (top left), pose multiple hypotheses (middle left), and construct models that explain how a hypothesis could have actually caused a particular phenomenon (the causal chain in the middle). Within these models, teams provide evidence in support of their model (top right). Teams in the Experimental group receive feedback from MILA–T, a metacognitive tutoring system. Here, the Interviewer is waiting to give feedback in the bottom left (as indicated by the light bulb), and the Guide (bottom right) is currently giving feedback in the bottom right.

In constructing these models, teams are asked to supply evidence in support of their claims. Evidence annotates the individual connections in the model. The box in the top right of Figure 1 shows the evidence that the team is supplying in support of the claim that the increased fertilizer quantity leads to an increase in phosphorus concentration. Justifying this connection demands three ideas: showing that fertilizer did increase, showing that phosphorus did increase, and showing that the two increases are causally linked. Evidential justifications are further annotated with categories: teams can choose from one of seven categories that describe their evidence derived from the epistemic cognition community and our prior research: logical explanations, expert information, non-expert information, direct observations, controlled experiments, simulation observations, and similar system observations (the first five derived from Goldin, Renken, Galyardt & Litkowski 2014, the remaining two added based on our activities). These evidence categories are associated with scores to reflect the desirability of different types of evidence in defending a hypothesis: it is, for example, preferable to rely on established scientific theories and observations from similar systems than to rely on logical explanations and novice information. These scores lead to the calculation of the evidence strength metrics described under data analysis later in this paper. (Joyner 2015 provides more details of MILA.)

## MILA–T: Metacognitive Tutoring

MILA is augmented with a metacognitive tutoring system that is the primary object of analysis for this paper. The metacognitive tutoring system, MILA–T, is comprised of five distinct agents: a Guide, a Critic, an Interviewer, a Mentor, and an Observer. Each of these tutors is defined by a functional role that a teacher typically plays in the classroom, tied in part to Grasha's (1996) model of teaching styles, and based partly on our observations of teaching and learning in science classrooms (Goel et al. 2013). This differentiates MILA–T from other metacognitive tutoring initiatives; whereas systems like MetaTutor define the functional roles of the agents with respect to the target skill (Azevedo et al. 2009, 2010), MILA–T defines the functional roles according to the pattern of interaction between the student and the agent.

The five tutors are classified into two broad categories: proactive and reactive. The proactive tutors (the Mentor, the Interviewer, and the Observer) continuously monitor the team's modeling process and intervene where necessary. The reactive tutors (the Guide and the Critic) wait for the team to solicit feedback. In this way, the proactive tutors mimic a teacher moving around the classroom, observing teams' progress, and intervening where necessary. The

reactive tutors mimic a teacher sitting at the front of the room waiting for teams to approach him or her for approval or guidance about how to proceed. These functional roles are further differentiated by the specific type of interactions that each tutor facilitates; the Critic, for example, critiques the current status of the team's explanation, while the Guide anticipates and answers the team's questions. Similarly, the Mentor monitors for mistakes or increasing aptitude, while the Interviewer waits for critical moments and asks students to reflect on their thought process.

MILA–T is defined as a metacognitive tutoring system, rather than a cognitive tutoring system, because the target of MILA–T is students' internal inquiry-driven modeling process. Thus, MILA–T is metacognitive in two ways: first, it itself reasons over students' thought processes, meaning that MILA–T thinks about students' thinking; and second, the skill it attempts to teach students is metacognitive. Inquiry-driven modeling is a metacognitive skill that operates on the learner's current understanding of a system, and thus the target of inquiry-driven modeling is the learner's own knowledge and understanding, meeting the common definition of a metacognitive skill (e.g., Veenman, Hout-Wolters & Aflerbach 2006).

The tutors of MILA–T take a three-prong approach to teaching metacognition: emphasize, instruct, and demonstrate. First, the tutors anticipate that students will deemphasize metacognition, and thus anticipate questions that students might ask; they then take those opportunities to turn students' attention toward the metacognitive skill. For example, the Guide anticipates students may ask what the right answer to the system is, and reacts to that question by describing to students how the "right" answer in science is an explanation they construct rather than an answer they receive. Second, the tutors attempt to explicitly instruct metacognition in their feedback. The Critic, for example, gives students feedback on what kind of evidence they rely on in their explanations, but augments this feedback with notes on why certain kinds of evidence are considered preferable and how one ought to evaluate an argument grounded in certain types of evidence. Third, the tutoring system, especially the Mentor and the Interviewer, attempt to demonstrate proper metacognition to the students. The Interviewer, for example, will react to certain critical decisions that students make by asking students to explain the reasoning that led to their decision, and then respond with an example of the reasoning she would have used to arrive at the same decision, thus demonstrating the desirable metacognitive process. (Joyner 2015 provides more details of MILA-T.)

## Experimental Design

The experiment with MILA-T was conducted with two middle school science teachers together teaching ten total classes. Participation in the experiment took place during nine consecutive regular school days, with each class participating for 45 minutes per day. The first and last days of this nine-day unit were spent on content and attitude testing; the third and sixth days were spent on laboratory exercises. The remaining five days of the unit were spent interacting with the exploratory learning environment MILA in teams of two or three. Throughout the first four days of interaction with MILA, teams of students were asked to develop an explanation of a massive fish kill that occurred in a nearby lake a few years earlier; this is dubbed the "Learning" project because it is during this time that teams are learning the skills associated with scientific modeling and inquiry. On the fifth day of interaction with MILA (the eighth day of the unit overall), teams are asked to develop an explanation of the record-high temperatures taking place in Atlanta over the past decade; this is dubbed the "Transfer" project because students are transferring the skills they learned to a new phenomenon. At the conclusion of each project, students submit their final explanation of the phenomenon. Thus, each team submits two models during the unit: a model of the fish kill and a model of Atlanta's high temperatures.

The controlled variable in this study is access to MILA–T. Both the Control and Experimental groups interact with MILA and complete the nine-day curriculum described above. The Control group never sees MILA–T. The Experimental group receives MILA–T during the Learning project online; MILA–T is disabled during the Transfer project. In this way, we may analyze whether students improve in their inquiry-driven modeling while receiving feedback from MILA–T by comparing the Control and Experimental during the Learning project, and we may also analyze whether any improvements persisted after the feedback from MILA–T was disabled by comparing the groups during the Transfer project.

Entire classes were assigned to either the Control or Experimental group in order to prevent Control group teams from being aware of the existence of MILA–T. Given the significant differences in teaching style between the two teachers, classes were assigned to one group or the other within the teacher; thus, each teacher taught three classes in the Experimental group and two classes in the Control group. As a result, 84 total teams completed the Learning project with 50 teams in the Experimental group and 34 teams in the Control group. 81 teams completed the Transfer project with 47 teams in the Experimental group and 34 teams in the Control group. Teachers assigned students to teams without any direction from the researchers. Researchers were present in the classroom to provide technical support, but avoided interacting with students on the project itself. Teachers taught identical material to classes in the Control and Experimental groups with no direction from the researchers to interact differently with the two groups of classes.

## Data Analysis

This data analysis focuses on the models that teams of students constructed to explain the two phenomena; other analyses have examined the impact on students' dispositions toward science (Joyner & Goel 2014), process of model

construction (Joyner & Goel 2015), and content knowledge (Joyner 2015), As described previously, while constructing models in MILA, teams annotate their models with evidential justifications for their explanation. These evidential justifications are the primary target for analysis here: how well do teams justify their explanations? To answer this, we performed two analyses: coded evidence analysis and quantitative model analysis.

### Coded Evidence Analysis

1301 total pieces of evidence were supplied in support of the 165 models. First, a subset of these pieces of evidence was randomly drawn and subjected to grounded analysis. Notes were taken on whether each piece of evidence was acceptable as a justification for the explanation, and if not, for what reason the piece of evidence was unacceptable. These notes were then processed into a coding scheme with seven categories: Acceptable, Miscategorized, Redundant, Gibberish, Irrelevant, Insufficient, and Not Evidence. The first two categories are generally noted as 'Acceptable' evidence (as Miscategorized evidence still supports the explanation, but simply is annotated with the wrong category), and the last five categories are noted as 'Unacceptable' evidence.

After establishing this coding scheme, all 1301 pieces of evidence were run through three rounds of coding by a single rater with three weeks between coding sessions. Intrarater reliability was established as very good between every pair of rounds of coding (Cohen's Kappa of 0.890 between the two most similar coding rounds). The results of this coding process were then tested using a $X^2$ analysis to examine whether the results were different between the Control and Experimental conditions. This was performed separately on the results of the Learning and Transfer projects.

### Model Analysis

Given the results of that evidence coding process, five metrics were calculated for each model. Each model's "total evidence strength" was calculated by summing the strengths of all pieces of evidence supplied in support of the model; strengths were assigned on a scale of 1 (weak) to 3 (strong) based on the category given to the evidence by the student. Each model's "average model strength" was calculated by dividing the total strength by the size of the model to analyze how well the team justified each individual claim of the model. Each model's "average evidence strength" was calculated by dividing the total strength of the model by the number of pieces of evidence to analyze the strength of the individual pieces of evidence supplied by the team. Each model's "total evidence" was calculated simply by counting the number of pieces of evidence without regard to strength. Each model's "model complexity" was calculated by counting the number of nodes and edges in the model.

All of these metrics were calculated with the evidence that resulted from the coded evidence analysis. Any piece of evidence that was coded as Unacceptable were not counted. Any pieces of evidence that were coded as Miscategorized were scored with the category assigned to them during the evidence coding process. These metrics were then analyzed using a multivariate analysis of variance to determine whether any differences existed in any of the metrics based on the experimental condition.

## Results

Both these analyses demonstrate the same conclusion: teams in the Experimental group outperformed teams in the Control group in both the Learning and Transfer projects.

### Coded Evidence Analysis Results

In order to examine the difference between the Control and Experimental groups as a result of this evidence coding process, a $X^2$ analysis was performed to determine whether the distributions of two samples were identical. $X^2$ analysis of the results of the coding process for the Learning project demonstrated a statistically significant difference between the Control and Experimental groups ($\chi^2 = 52.423$, $p \approx 0.0$). The Experimental group outperformed the Control group: 59.00% of the pieces of evidence written by the Control group teams were coded as Acceptable, while 72.83% of the pieces of evidence written by the Experimental group teams received that positive designation. This improvement results in significantly lower proportions of the Experimental group's evidence falling into several negative categories, as documented in Table 1.

Table 1: 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. Column labels show the seven categories assigned in coding: Acceptable, Miscategorized, Redundant, Gibberish, Irrelevant, Insufficient, Not Evidence.

| | A | M | G | I | N | R | S |
|---|---|---|---|---|---|---|---|
| Control | 59.0% | 12.6% | 5.8% | 8.4% | 6.1% | 5.3% | 2.6% |
| Expected | 271 | 27 | 39 | 58 | 28 | 25 | 12 |
| Observed | 335 | 9 | 25 | 58 | 9 | 21 | 3 |
| Experi-mental | 72.8% | 12.6% | 2.0% | 5.4% | 2.0% | 4.5% | 0.6% |

$X^2$ analysis of the results of the coding process for the Transfer project also demonstrated a statistically significant difference between the Control and Experimental groups ($\chi^2 = 42.720$, $p \approx 0.0$). The Experimental group again outperformed the Control group: 55.41% of the pieces of evidence written by the Control group teams were coded as Acceptable, while 67.91% of the pieces of evidence written by the Experimental group teams received that positive designation. This improvement results significantly lower proportions of the Experimental group's evidence falling into several negative categories, as documented by Table 2.

Table 2: 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 | 55.4% | 11.7% | 5.2% | 8.2% | 7.8% | 10.9% | 0.9% |
| Expected | 193 | 18 | 29 | 41 | 27 | 38 | 3 |
| Observed | 237 | 10 | 17 | 43 | 18 | 16 | 8 |
| Experi-mental | 67.9% | 12.3% | 2.9% | 4.9% | 5.2% | 4.6% | 2.3% |

Between the Learning and the Transfer project, both the Control and the Experimental groups experienced a slight but statistically significant decrease in the overall acceptability of their evidence ($\chi^2 = 15.62$, $p < 0.05$ for the Control group; $\chi^2 = 37.94$, $p \approx 0.0$ for the Experimental group), likely based on the reduced time available for the Transfer project. The Experimental group experienced a larger decrease than the Control group, however, suggesting that the tutoring system had played a role in improving the Experimental group's models during the Learning project. Despite this larger decrease in evidence acceptability, however, the Experimental group teams still outperformed the Control group teams in the Transfer project.

In terms of the learning goals for the project, these results show that teams in the Experimental group have an improved ability to use evidence in support of their arguments and explanations compared to teams in the Control group. It is thus reasonable to say that participation with MILA–T during the unit improved teams' inquiry-driven modeling by improving their ability to use evidence in support of their claims.

### Model Analysis Results

The results of this evidence coding process were used to score teams' models according to the five metrics described previously. A multivariate analysis of variance was then performed for the Learning and Transfer projects to discern whether there was a difference in teams' performance along these metrics based on participation in the Experimental condition. If the multivariate analysis of variance revealed an effect of the condition, a follow-up univariate analysis was conducted on each of the five variables to discern which variables were impacted.

During the Learning project, there existed a statistically significant effect of the experimental condition ($F = 3.3$, $p < 0.01$). Follow-up univariate analysis showed a statistically significant influence of Condition on three variables during the Learning project: total model strength ($F = 10.9$, $p < 0.01$), average model strength ($F = 7.5$, $p < 0.01$), and total evidence ($F = 5.9$, $p < 0.05$). This means that teams in the Experimental group constructed better-justified models in terms of both the overall strength of the justification and the average strength of the justification supplied for each individual claim in the model. The improvement was approximately 50% across each of the three metrics, meaning that the models that the models produced by teams in the Experimental group were approximately 50% stronger (as measured by these metrics) than models produced by teams in the Control group. Thus, during the Learning project, teams in the Experimental group constructed better-justified models than teams in the Control group.

During the Transfer project, there again existed a statistically significant effect of the experimental condition ($F = 4.2$, $p < 0.01$). Follow-up univariate analysis showed a statistically significant influence of Condition on all five variables during the Transfer project: total model strength ($F = 12.0$, $p < 0.001$), average model strength ($F = 6.0$, $p < 0.05$), total evidence ($F = 4.2$, $p < 0.05$), average evidence strength ($F = 12.2$, $p < 0.001$), and model complexity ($F = 4.9$, $p < 0.05$). This means that the Experimental group teams outperformed the Control group teams in several ways: they supplied more evidence, the individual pieces of evidence they supplied were stronger, and the combination of these strengths led to stronger justifications of each claim in their models and their models as a whole. This improvement was approximately 70% across each of these metrics; thus, teams in the Experimental group produced models that were approximately 70% stronger than models produced by teams in the Control group.

In terms of the learning goals for this project, these results show that the previously documented increased ability to use evidence in support of arguments and explanations has increased the strength of the final arguments that these teams produce. It is thus reasonable to say that participation with MILA–T during the unit improved teams' inquiry-driven modeling as demonstrated by improvements seen in the strength of the models that these teams produced.

These analyses were also run on the raw evidence prior to the coding process, and while the results differed, the Experimental group teams outperformed the Control group teams in similar ways.

## Conclusions

The above results indicate that the teams who received the metacognitive tutoring system MILA–T outperformed teams that did not receive the tutoring system during the same activity. More significantly, however, this superior performance also carried through to a new activity where MILA–T was no longer available, demonstrating that the feedback that teams received from MILA–T was internalized and transferred to a new task. This suggests that access to MILA–T improved teams' inquiry-driven modeling, and thus improved the quality of the models the teams generated. These improvements were seen both in the individual evidence that teams provided and in the total strength of the justifications for the teams' models as a whole.

However, it is important to note that some of these improvements are difficult to attribute directly and solely to the metacognitive tutoring system. During both the Learning and Transfer projects, teams in the Experimental group demonstrated a decreased propensity to supply unacceptable evidence. The acceptability of evidence is determined solely by the text that the team provides, and MILA–T is unable to read this text; it can only give feedback on the categories that teams choose for their evidence. Thus, despite MILA–T's inability to give feedback on the actual text of the justifications that teams provide, the text nonetheless improved.

So how can one explain this improvement among teams receiving MILA–T without MILA–T giving direct feedback on the quality of this text? One explanation is that MILA–T supplied information on what makes a good justification, and this information was internalized by teams even without receiving feedback on their own present justifications; this explanation, however, relies on very significant improvement without any targeted feedback. Another explanation is that teams in the experimental group felt observed given the visual presence of the tutoring system and thus were more likely to engage more earnestly, leading to better justifications; this explanation relies on teams carrying over these good habits into the Transfer project even after the tutoring system has been disabled.

The third, and we posit the most likely, explanation is corroborated by teacher feedback. In exit interviews, teachers in the experiment reflected that in Experimental classrooms, they were not needed as often to answer basic low-level questions because the tutoring system took care of these simple feedback opportunities; they, then, were able to focus on high-level and complex feedback that went beyond the scope of what the tutoring system could provide. This reflects aninterplay between individual- or team-level tutoring systems and classrooms as a whole; in many ways, intelligent tutoring systems can be seen as approaches to offloading responsibilities from the teacher to allow the teacher to more effectively direct their time toward feedback that only a human can provide.

This also lends evidence in support of our proposal for organizing metacognitive tutoring for inquiry-driven modeling around the functional roles of teachers in science classrooms. We plan to conduct additional research to examine whether the more and the better the metacognitive tutors can imitate the functional roles of teachers, (1) the higher is the efficacy of scaffolding learning about inquiry-driven modeling, and (2) larger is the offloading of responsibilities from the science teacher to the metacognitive tutors.

## References


Azevedo, R., Johnson, A., Chauncey, A., & Burkett, C. (2010). Self-regulated learning with MetaTutor: Advancing the science of learning with MetaCognitive tools. In M. Khine & I. Saleh (Eds.), *New science of learning: Computers, cognition, and collaboration in education* (pp. 225-247). Amsterdam: Springer.

Azevedo, R., Witherspoon, A., Chauncey, A., Burkett, C., & Fike, A. (2009). MetaTutor: A MetaCognitive tool for enhancing self-regulated learning. In *Procs. AAAI Fall Symposium on Cognitive and Metacognitive Educational Systems*.

Clement, J. (2008). *Creative Model Construction in Scientists and Students: The Role of Imagery, Analogy, and Mental Simulation*. Dordrecht: Springer.

Darden, L. (1998). Anomaly-driven theory redesign: computational philosophy of science experiments. In T. W. Bynum, & J. Moor (Eds.), *The Digital Phoenix: How Computers are Changing Philosophy* (pp. 62–78). Oxford: Blackwell.

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.

Goel, A., Rugaber, S., Joyner, D., 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 *International Handbook on Meta-Cognition and Self-Regulated Learning,* R. Azevedo & V. Aleven (editors), pp. 545-560, Springer.

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*. Springer International Publishing.

Grasha, A. (1996). *Teaching with Style*. Pittsburgh: Alliance Publishers.

Joyner, D. (2015). *Metacognitive Tutoring for Inquiry-Driven Modeling.* Ph.D. Dissertation, School of Interactive Computing, Georgia Institute of Technology, Atlanta, Georgia.

Joyner, D. & 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.

Joyner, D. & 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.

National Research Council. (1996). *National Science Education Standards.* Washington, DC: National Academy Press.

Nersessian, N. (2008). *Creating Scientific Concepts*. Cambridge, MA: MIT Press.

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).

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 *Procs. International Conference on Intelligent Tutoring Systems*. Berlin: Springer.

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.

Veenman, M., Van Hout-Wolters, B., & Afflerbach, P. (2006). Metacognition and learning: conceptual and methodological considerations. *Metacognition and Learning*, *1*, 3–14.

White, B. & Frederiksen, J. (1998). Inquiry, modeling, and metacognition: Making science accessible to all students. *Cognition and Instruction, 16*(1). 3–118.