Enhancing Scientific Reasoning and Discussion with Conversational Agents

Gregory Dyke, David Adamson, Iris Howley, and Carolyn Penstein Rosé, *Member, IEEE*

Abstract—This paper investigates the use of conversational agents to scaffold online collaborative learning discussions through an approach called academically productive talk (APT). In contrast to past work on dynamic support for collaborative learning, which has involved using agents to elevate the conceptual depth of collaborative discussion by leading students in groups through directed lines of reasoning, this APT-based approach lets students follow their own lines of reasoning and promotes productive practices such as explanation of reasoning and refinement of ideas. Two forms of support are contrasted, namely, Revoicing support and Feedback support. The study provides evidence that Revoicing support resulted in significantly more intensive reasoning exchange between students in the chat and significantly more learning during the chat than when that form of support was absent. Another form of support, namely, Feedback support increased expression of reasoning while marginally decreasing the intensity of the interaction between students and did not affect learning.

Index Terms—Collaborative learning, intelligent agents, psychology

1 Introduction

A large body of work has shown that certain forms of classroom interaction, termed accountable talk, or academically productive talk (APT), are beneficial for learning with understanding in subjects such as math and science [17]. In this paper, we explore how we can achieve some of the benefits of this form of learning support within small online groups engaged in learning scientific content supported by technology.

In prior work using intelligent conversational agents to support collaborative learning, the agents have provided social support, affording the agents a more credible social standing in the group and helping to diffuse tension and create a productive learning environment [14]. Furthermore, they have provided conceptual support, designed to elicit more depth by leading students through directed lines of reasoning, referred to as knowledge construction dialogues (KCDs) [2], [13], [14]. KCDs have been shown to increase learning gains in Science [19], Math [12], and Engineering [13], particularly in situations where the conversational agents also provide social support [2], [14]. However, the necessity of designing them statically, with a predefined line of reasoning in mind, both makes them hard to adapt to new subject material and does not fully exploit the benefits of collaborative learners following their own spontaneous lines of reasoning.

We have, therefore, drawn on and integrated extensive work related to the support of classroom discourse, to investigate the use by conversational agents of facilitation moves that promote academically productive talk [17]. The aim of APT facilitation moves is to increase the amount of displayed reasoning and *transactivity* [5], which is the extent that learners build on the ideas of others. The extent to which learners build on each other's contributions rather than simply talking about their own ideas can be thought of as a desirable *intensity* in the interaction. This is achieved by dynamically reacting to student discussions and directing them to listen to and respond to each other in constructive ways. Furthermore, as APT refers both to learners' social positioning with respect to each other and their conceptual positioning with respect to knowledge, it provides us with a theoretical framework to better integrate the social and conceptual support aspects of conversational agents in a generalizable and scalable way.

In this paper, we first discuss the theoretical foundation for our work from the classroom discourse and computer supported collaborative learning communities. We then describe a new architecture for enabling the flexible development of orchestrated, APT-based dynamic collaborative learning support. Finally, we describe a classroom study involving students from seven ninth grade biology classrooms that provides evidence of significant positive effect of multiple forms of APT-based support.

2 Theoretical Framework

The theoretical foundation for the work reported in this paper comes from three areas. Specifically, first we draw from the literature on academically productive talk. Next, we draw from the literature on scripted collaboration from the computer supported collaborative learning community. Finally, we draw from the recent literature on dynamic support for collaborative learning.

- *G. Dyke is with the Université de Lyon, ENS Lyon, Lyon, France. E-mail: gregory.dyke@ens-lyon.fr.*
- *D. Adamson is with the Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA 15206. E-mail: dadamson@cs.cmu.edu.*
- *I. Howley is with the Human Computer Interaction Institute, Carnegie Mellon University, Pittsburgh, PA 15206. E-mail: iris@cmu.edu.*
- *C.P. Rosé is with the Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA 15206. E-mail: cprose@cs.cmu.edu.*

Manuscript received 7 Jan. 2013; revised 24 Apr. 2013; accepted 17 June 2013; published online 28 June 2013.
For information on obtaining reprints of this article, please send e-mail to: lt@computer.org, and reference IEEECS Log Number TLT-2013-01-0003.
Digital Object Identifier no. 10.1109/TLT.2013.25.

2.1 Academically Productive Talk

Academically Productive Talk has grown out of frameworks that emphasize the importance of social interaction in the development of mental processes, and has developed in parallel to similar ideas from the computer-supported collaborative learning community. Michaels et al. [17] describe some of the core dialogic practices of academically productive talk along three broad dimensions:

1. Students should be accountable to the learning community, listening to the contributions of others and building on them to form their own.
2. Students should be accountable to accepted standards of reasoning, emphasizing logical connections, and drawing reasonable conclusions.
3. Students should be accountable to knowledge, making arguments that are based explicitly on facts, written texts or other public information.

Such practices are unfamiliar in many classrooms. Not only must they be introduced to students, it is also necessary to provide teachers with the means to scaffold and support these interaction forms. Drawing on over 15 years of observation and study, Michaels et al. propose a number of core "moves" that teachers can draw upon to encourage the development of academically productive classroom discussion. This set of moves includes:

1. Revoicing a student statement: "So let me see if I've got your thinking right. You're saying XXX?"
2. Asking students to apply their own reasoning to someone else's reasoning: "Do you agree or disagree, and why?"
3. Prompting students for participation: "Would someone like to add on?"
4. Asking students to explicate their reasoning: "Why do you think that?"

The teacher's facilitation plays a key role in encouraging students to display their reasoning and build on each other's reasoning, and, importantly, does not lead to a teacher-centered discussion. Instead, the teacher uses academically productive talk to hold students accountable for their own knowledge and reasoning, and to remind them to hold themselves and each other accountable likewise. In studies where teachers used approaches like academically productive talk, students have shown steep changes in achievement on standardized math scores, transfer to reading test scores, and retention of transfer for up to 3 years [1].

2.2 Script-Based Support for Collaboration

The computer supported collaborative learning community shares many of the same values related to desired conversational practices in student group discussions. For example, externalizations of reasoning and connections to prior reasoning, as captured by the Transactivity construct [5], have been shown to be positively correlated with learning in collaborative environments [3]. This encouragement toward connected displays of reasoning is quite similar to APT's aims of accountability to group and individual knowledge and reasoning. What is different is that a teacher is normally not present to support practices within an online collaborative setting. Thus, it is necessary to design environments with affordances that play the same role, to whatever extent is possible. The most popular approach to providing such affordances in the past decade has been that of script-based collaboration [11].

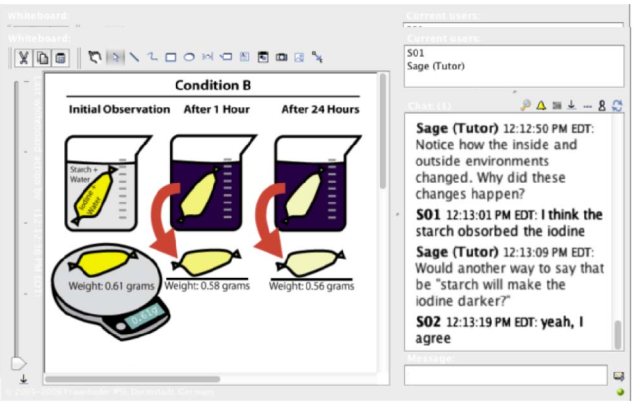
A collaboration script may describe any of a wide range of features of collaborative activities, including its tasks, timing, the distribution of roles, and the methods and desired patterns of interaction between the participants. Scripts can be classified as either macroscripts or microscripts [7]. Macroscripts are pedagogical models that describe coarse-grained features of a collaborative setting, which sequence and structure each phase of a group's activities by attributing roles and tasks to foster collaboration. Microscripts, in contrast, are models of dialogue and argumentation that are embedded in the environment, and are intended to be adopted and progressively internalized by the participants. Examples of macroscripts include the classic Jigsaw activity, as well as more tailored approaches like ArgueGraph and ConceptGrid [10]. Microscripting can be implemented by offering prompts or hints to the user to guide their contributions [21], which may depend on the current phase of the macroscript. Traditional collaboration scripts such as these can support both conversational and reasoning practices, but fall short of providing the active facilitation described by the APT literature.

2.3 Dynamically Scripted Support

Early approaches to scripted collaboration, as described above, have been static, offering the same script or supports for every group in every context. Such nonadaptive approaches can lead to overscripting [6], or to the interference between different types of scripts [22]. A more dynamic approach that triggers microscripted supports or the appropriate phases of macroscripts in response to the automatic analysis of participant activity [20] would be preferable. This analysis could occur at a macrodiscourse level, following the state of the activity as a whole, or it could be based on the classification of individual user contributions. Further, the benefits of fading the support over time [21] could be more fully realized, as the timing and degree of such fading could be tuned to the group's level of internalization. The collaborative tutoring agents using the Basilica agent architecture described by Kumar and Rosé [15] were among the first to implement dynamic scripting in a CSCL environment and demonstrate gains over otherwise equivalent static support. More recent work by Baghaei et al. [4] and Diziol et al. [8] similarly show that while the field of adaptive support for collaborative environments is still in its infancy, such supports can produce effects on student interaction and learning.

3 Conversational Agent Architecture

Just as human teachers orchestrate elements of collaborative learning in their classrooms, a conversational agent-as-facilitator must concurrently manage several differently scoped supports and behaviors, including collaborative macroscripts, social cues, and adaptive facilitation. The Basilica agent architecture [15] was among the first to implement support for the modular composition of tutor



S01
Sage (Tutor)

Sage (Tutor) 12:12:50 PM EDT: Notice how the inside and outside environments changed. Why did these changes happen?

S01 12:13:01 PM EDT: **I think the starch obsorbed the iodine**

Sage (Tutor) 12:13:09 PM EDT: Would another way to say that be "starch will make the iodine darker?"

S02 12:13:19 PM EDT: **yeah, I agree**

Fig. 1. The collaborative environment used in the studies described. At left, items for discussion are displayed. At right, students interact with each other and the agent through text chat.

behaviors within a CSCL environment. The open-source Bazaar architecture described in this paper builds upon Basilica to address the challenge of orchestration, replacing the former's ad hoc interbehavior communication with a framework specifically designed for coordinating multidimensional collaborative agents.

In the next two sections, we describe the behavioral components implemented to provide APT facilitation for this study. In Section 3.3, we describe the remaining components of our agent, and briefly outline the way Bazaar explicitly addresses orchestration between behaviors.

3.1 Automatic Revoicing

One of the forms of support evaluated in this paper is a Bazaar component that performs a form of APT referred to as Revoicing. The agent compares student statements against a list of conceptually correct statements developed with teachers. In the study described in this paper, 64 such statements were developed and validated against pilot data. For each student turn, we calculate a measure of "bag of synonyms" cosine similarity against each expert statement, based on the method described by Fernando and Stevenson [9]. If this similarity value exceeds a set threshold, we consider the student's turn to be a possible paraphrase of the matched statement, and thus "revoicable." If the matched entry has not triggered a revoice before, the Revoicing component responds by offering it as a paraphrase of the student's turn, for example "So what I hear you saying is XXX. Is that right?" An example of this behavior is displayed in context in Fig. 1, with the text of the interaction contained in Table 1. All occurrences of revoicable turns are logged for later process analysis, independent of the agent's performance of revoicing moves.

3.2 Academically Productive Feedback

Another manipulation implemented using Bazaar and evaluated in this study is a component that provides positive feedback for APT. Student input is matched against a list of expressions indicating the performance of transactive reasoning and APT moves, based on the descriptions by Michaels et al. [17], including explanation, challenge, revoicing, and requests for others to provide each of the same. If a student statement matches, the agent publicly praises the student's move, and (when appropriate) encourages the other students to respond. All students who participated in the study reported in this paper received instruction about APT in the form of a cartoon illustrating the discussion moves prior to the online collaborative activity. Rather than perform APT-based facilitation itself, as the Revoicing behavior did, the Feedback behavior was meant to indirectly support the prevalence of APT in the discussions by rewarding students for taking this facilitation role.

TABLE 1
Agent Facilitation Excerpt

T *script*	**What changes do you think you'll see in this condition? Agree on a prediction for what you'll observe over time, *and* an explanation that everyone understands.**
S01	yeah it creaced after time went on and more glucose moved through the strip
S01	i think after time goes by glucose wont enter ther distilled water
T *feedback*	**Thanks for starting a prediction, S01.**
S02	maybe you'll see that the color of the strip changes because the glucose passes through the membrane of the cell therefore decreasing the amount outside and leaving little to none for the indicator to pick up
T *revoice*	**Maybe you could state that as "the glucose can pass through the cell membrane."**
S02	yeah you could say that
T *social*	**I wonder what S03 has to say.**
S01	would that be our explanation ?
S02	lets see what S03 has to say first then we'll get an explaination

3.3 Orchestrating Agent Behavior

The revoicing prompts and APT feedback behaviors manipulated in this study were performed by the agent in tandem with other forms of support that were common across all conditions. One such support was a static collaborative macroscript that structured the overall timing and flow of the activity, providing prompts for each problem set and updating the figures displayed on the whiteboard at predetermined intervals. Another shared component inserted dymanically triggered social prompts designed to encourage participation and group cohesion, as employed in earlier work by Kumar et al. [14].

The orchestration of multiple supports is enabled by the Bazaar architecture. The user-facing actions proposed by each behavioral component are delivered to the architecture's Output Coordinator, which periodically selects and enacts these proposals from a priority queue. Accepted proposals can install "advisors" that temporarily influence the priority of future proposals, allowing a component to "hold the floor" or promote other types of followup actions.

This facet of the Bazaar architecture is illustrated in Fig. 2—in the example configuration shown, when the revoice move is accepted, it installs an advisor that blocks additional proposals long enough for the students to

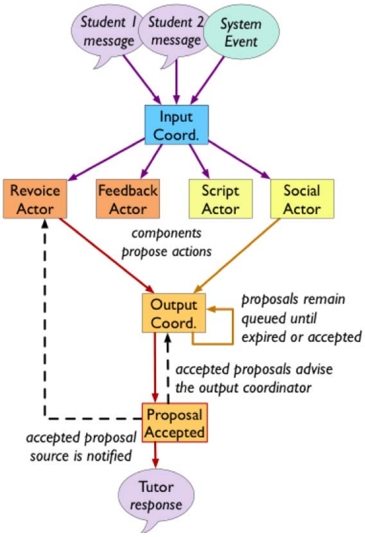

Student 1 message
Student 2 message
System Event
Input Coord.
Revoice Actor
Feedback Actor
Script Actor
Social Actor
components propose actions
Output Coord.
proposals remain queued until expired or accepted
accepted proposals advise the output coordinator
Proposal Accepted
accepted proposal source is notified
Tutor response

Fig. 2. Overview of Bazaar component architecture.

process the move and respond. The Social proposal is meant as an immediate response to a particular student turn, and thus was defined with a short timeout and a low priority—it will likely expire before the Revoice advisor releases the floor. Similarly, the macroscript component may propose the next segment of timed dialogue while the Revoice advisor is still active—but (in this particular configuration) as the script move's precise timing is less important, it is defined with a longer timeout, and will wait for the revoicing exchange to finish before progressing with the script. This flexible coordination of independent conversational behaviors allows the quick composition of dynamic experimental manipulations with standard collaborative supports.

4 Method

In accordance with the literature on APT used as a classroom facilitation technique, in this study we test the hypothesis that appropriate APT support in a computer-supported collaborative learning setting will:

- Intensify the exchange of reasoning between students during the collaborative activity;
- Increase learning during the activity; and
- Better prepare students for subsequent learning outside of the small group environment.

4.1 Instructional Content and Study Procedure

4.1.1 Participants

This study was conducted in seven ninth grade biology classes of an urban school district. The classes were distributed across two teachers (with, respectively, 3 and 4 classes) for a total of 78 consenting students, who were randomly assigned to groups of 3. Groups were randomly assigned to conditions.

4.1.2 Experimental Manipulation

This study was run as a 2×2 between subjects factorial design in which the APT agents provided some behaviors in common across conditions, but other behaviors were manipulated experimentally. Across all conditions, the agent provided the same macro level support by guiding the students through the activity using the same phases introduced in such a way as to control for time on task. It was the microscripting behaviors that were manipulated experimentally to create the four conditions of the 2×2 factorial design. The first variable for manipulation was the presence or absence of the Revoicing behavior described in Section 3.1. The second variable was the presence or absence of the APT Feedback behavior described in Section 3.2.

In addition, in each classroom session, a group was provided with "Wizard of Oz" support in which a human experimenter performed both revoicing and feedback. We did this to assess whether any deficiency in positive effect of either factor might be due to technical failure rather than poor design. Results in the Wizard conditions on all measures were always within the same range as in the fully automatic support conditions.

4.1.3 Learning Content

The study was carried out during a module introducing the concepts of selective permeability, diffusion, osmosis, and equilibrium. In this module, students observe that glucose, water, and iodine molecules all diffuse through dialysis tubing while starch molecules do not. The activity naturally lends itself to observing a variety of distinct cell models involving dialysis tubing containing an *inside environment* immersed in a beaker containing the *outside environment*. In each, a choice must be made for which liquid will be placed outside and which liquid will be placed inside. Four were used in the study:

1. Model A includes a starch suspension inside dialysis tubing and iodine solution outside (the iodine serves as an indicator for starch).
2. Model B is the opposite of A, having the iodine solution within the dialysis tubing and the starch suspension outside.
3. Model C includes a glucose solution on the inside of the tubing and distilled water outside.
4. Model D is the opposite of C, with distilled water in dialysis tubing and glucose solution outside.

In the case of cell models A and B, movement of the starch suspension and iodine solution can be detected through a change in color of the inside or outside environment. In the other two models, indicator strips that change color in the presence of glucose can detect whether the glucose solution has mixed with the water.

The collaborative task content, the macroscripts that supported it, and the list of key concepts used for revoicing were all developed iteratively with feedback from teachers and content experts.

4.1.4 Study Procedure

The study was conducted over three phases, which occurred as single class periods over two school days.

The first phase ("day 1") involved the teachers running a lab as a demonstration of building a cell model with their students as they would normally with cell model A, the condition of starch suspension inside dialysis tubing and iodine solution outside. The students observe the cell model as it is constructed and then 24 hours later. The students took a pretest at the end of this first phase.

The second phase ("day 2") was centered around a 20 minute collaborative computer-mediated activity during which the experimental manipulation took place. The students did the activity in groups of three students, scaffolded by academically productive talk agents. Students within classes were randomly assigned to groups and then groups to conditions. This activity was introduced by a cartoon depicting the use of APT, a reminder of the results of the previous day (with cell model A) and an introduction to the "new" information: glucose and glucose test strips. The conversational agent led the students through two new models: cell models B and C.

For each of these models, the agent showed the outcomes after 1 and 24 hours in terms of the colors inside and outside (indicating whether starch and glucose had diffused in or out) and the weight of the tubing (indicating whether water had traveled). For each observation, the agent asked the students to come up with an explanation. The agent then presented the students with cell model D, glucose outside and water inside (the opposite of model C) and asked the students to collaboratively come up with a prediction for what they would observe, and an explanation for their prediction. They were instructed to write down their prediction and explanation when they were in agreement and were informed that there would be prizes for the best explanations. To assist them in this activity, students were given a worksheet summarizing the setup for each condition and providing space to write down their prediction and explanation for cell model D. At the end of this second phase, the students took the Postactivity test.

The computer activity was intended to equip the students with enough empirical data and attempts at reasoning to prepare them for the third phase ("day 3"), a full class APT discussion with their teacher, during which they would reconcile their different understandings and explanations. At the end of this discussion, they took a Postdiscussion Test.

4.2 Measurement

Domain knowledge was measured at three time points using a paper based test. Each of the three tests (Pretest, Postactivity Test, Postdiscussion Test) followed a similar format: a multiple choice question, a fill in the blank question, and what we refer to as a *concept cartoon*, which displayed a scenario that a student was required to generate an explanation about. In particular, the idea of the concept cartoon is to present a contextualized situation with three statements which can all be true given certain assumptions. Respondents are asked to pick the statement they are most in agreement with and to explain why they agree. As an example, the text from the postactivity test's concept cartoon question is given below.

> *"We fill the same sort of dialysis tubing from our earlier experiments with pure water, and place it in a pitcher of Kool-Aid. Which of the statements about the next 24 hours do you agree with the most? Explain your reasoning.*
>
> *A: The liquid inside the tubing will taste sweeter than the liquid on the outside.*
>
> *B: The liquid inside the membrane won't taste sweet at all.*
>
> *C: The liquid inside will taste just as sweet as the liquid on the outside."*

Each test covered the same knowledge but used different scenarios. The knowledge to be covered by each test was established in coordination with the teachers, with teacher trainers who identified common misconceptions, and with test results from a pilot run the previous year.

Each of the concept cartoon explanations was graded along four dimensions: the number of science terms used properly in a way that demonstrates understanding (e.g., "diffuse through the membrane" as opposed to "went through the bag") and the degree to which their explanation addressed each of the three learning objectives of the activity: concepts of the scientific method, movement of molecules, and the behavior of semipermeable membranes. Thus, for each test, we compute a per-objective score for each learning objective, and a total score, which is the sum across learning objectives. After an initial round of consensus coding by two graders on a sample of each test to establish a coding manual, all the tests were graded by a single grader.

4.3 Process Analysis

The goal of the intervention was to engage students in a more intensive exchange of explanations, which we referred to above as revoicable assertions. By more intensive, we do not mean that students utter more explanations per se, but that the expanations they utter are directed toward and building on those of their partner students. Anecdotally, we observed that in some conversations, there were bursts of explanation behavior where this kind of intensive knowledge exchange was taking place. The purpose of our quantitative process analysis was to measure the extent to which this kind of bursty behavior was occurring within discussions as a result of the manipulation.

To accomplish this, the chat logs were segmented into 5 minute intervals such that one observation is extracted per student for each interval. In each observation, we counted the number of revoicable assertions contributed by the student and the number of revoicable assertions contributed by other group members. Conversations with more bursty behavior patterns should have a higher correlation between these two variables, which would signify that students are more active in the conversation when their partner students are also active.

Thus, for the process analysis, we evaluate the effect of condition on the correlation within time slices between occurrence of revoicable assertions of a student with those of the other students in the same group. For this analysis, we used a multilevel model to analyze the results to account for nonindependence between instances. We expect to see that the correlation is significantly higher in the condition with the intervention. We do the analysis separately for each of the two interventions, namely, the Revoicing agent and the Feedback Agent. Specifically, we used what is referred to as a *random intercept and slope model*, which allows estimating a separate latent regression line for a student's behavior in relation to that of their partner students within time slices. In this model, each student

TABLE 2
Summary of Results per Condition, Mean (Standard Deviation)

	Control	Feedback Only	Revoicing Only	Revoicing and Feedback
Pre Test	.15 (.07)	.09 (.08)	.11 (.06)	.11 (.08)
Post-Activity Test	.2 (.22)	.16 (.15)	.21 (.2)	.18 (.1)
Post Discussion Test	.25 (.23)	.18 (.11)	.28 (.25)	.23 (.17)

trajectory is characterized by a regression with latent slope and intercept.

To do this analysis, we used the Generalized Linear Latent and Mixed Models (GLLAMM) [18] add-on to STATA. The dependent measure was number of revoicable assertions by the student within the time slice. The independent variable was the number of revoicable assertions contributed by the other students in the group within the same time slice. The condition variable was added as a fixed effect, and as an interaction term with the independent variable. A significant interaction between condition and independent variable in this case would indicate a significant difference in correlation between a student's contribution of revoicable assertions and that of their partner students, which would be indicative of an intensification of the interaction between students. A significant difference in intercept between conditions would indicate that the intervention raised the average number of revoicable assertions within time slices.

4.4 Results

In this study, we have tested the hypothesis that offering dynamic microscripting support to computer supported collaborative learning groups in the style of APT facilitation will produce more learning during collaborative learning discussions, will enrich the interactions between students, and will also better prepare them for participation in a whole group, teacher lead discussion.

As mentioned above, two independent manipulations were used to operationalize APT facilitation in this study, namely, Revoicing and Feedback. To evaluate the hypothesis, we took three measurements of domain knowledge, and conducted a process analysis of the interaction. To measure learning, we offered a Pretest, Postactivity test, and Postdiscussion test. Learning specifically between Pretest and Postactivity test is learning during the experimental manipulation. To measure preparation for participation in the whole group discussion, we also evaluated learning between the Postactivity test and the Postdiscussion test.

Some data was incomplete due to students being absent from class on one of the three study session days. Our analysis is, therefore, based only on teams where all three students were present on all three study session days. Altogether, three groups were dropped from the analysis, each from a different condition, leaving us with a total of 69 students.

The results per condition are summarized in Table 2, where test scores are expressed as percentages of the total composite test score, i.e., 0.7 signifies that 70 percent of possible points on the rubric were achieved. In this section, we detail our analyses and findings.

As an additional methodological point, within the condition that included both Revoicing and Feedback, there was one team per class session for whom the intervention was performed by a human selecting prompts from a list, as mentioned earlier in the paper. We conducted all of our analyses both with these data points included and without, and in no case were the results different. In fact, the average test scores for the Wizard sessions in all cases was very slightly lower than the condition average. Thus, in all cases, we include those data points in the analysis presented here.

First, we verified that the students learned from the online activity. For this analysis, we treated the three tests as repeated measures and built an ANOVA model with Test as dependent measure, and Time Point included with Feedback, and Revoicing as independent variables. We also included all two-way interactions and the one three-way interaction term. The result was that there was a significant effect of time point $\mathrm{F}(2, 191) = 9.25, p < 0.0001$, demonstrating that learning took place during the online activity. There was no significant effect of any other variable, showing that there was learning in all conditions. In a student-t posthoc analysis, we found that the difference between pretest and the other two tests was significant, but that the difference between the Post-Test and the Postdiscussion test was not significant. However, the effect size (*cohen's d*) of the difference between Pretest and Postdiscussion test was larger than that between Pretest and Post-Test, i.e., 0.74 s.d. versus 0.45 s.d. using the pooled standard deviation (0.16).

As a more fine grained test of learning, we used instead of the total test score as the dependent variable, the per-learning-objective score for the three learning objectives, namely, scientific method, movement of molecules, and semipermeable membranes. Thus, we had three observations per student, one for each learning objective. For this analysis, we added an additional independent variable referred to as Objective as well as the interaction between this variable and the Time Point variable to test for differential learning across learning objectives. This analysis showed a more nuanced pattern. Specifically, we see a significant interaction between objective and Time Point, that shows that the significant gain for some learning objectives occurred during a different phase $\mathrm{F}(4, 410) = 3.2$, $p < 0.05$. In particular, there was no significant difference across test phases for the scientific method. There was a significant difference between Pretest and Post-Test on the concept of semipermeable membranes, but not between Post-test and Post-Discussion Test. As for movement of molecules, we see significant gains between Post-Test and Post-Discussion Test, but not between Pre-Test and Post-Test. Since we see differential learning across learning objectives, in subsequent analyses of learning, we retain the Objective variable in our analyses.

Next, we evaluated the effect of the experimental manipulation on learning. First, we confirmed that our random assignment was successful in assigning students to groups that were roughly equivalent with respect to prior knowledge. We did this by using an ANOVA, with Revoicing and Feedback as independent variables and per-learning-objective Pretest as the dependent variable.

We also included an interaction term for the interaction between Revoicing and Feedback. Finally, we included the Objective variable as a final independent variable, and the interaction between Objective and all other variables and interaction terms. There were no significant effect of either condition variable or the interaction on pretest score. Thus, students in all conditions began with about the same amount of prior knowledge. The pattern was the same when considering the Objective variable. Thus, prior knowledge was consistent across conditions for all learning objectives.

Then we tested the effect of the experimental manipulation on learning during the collaborative activity using an ANCOVA with the per-learning-objective Postactivity test variable as the dependent variable and per-learning-objective Pre-Test variable as a covariate. All of the independent variables and interaction terms were the same as in the previous analysis. In this analysis, we see a significant effect of the Revoicing Condition $\mathrm{F}(1, 170) = 5.06, p < 0.05$ effect size 0.34 s.d., and no interaction with Objective. There were no other significant main effects or interactions. There was no significant effect of the Feedback manipulation. And though there was no significant interaction effect, our observation was that in the condition where students received both manipulations, there was some evidence that the interventions interfered with each other. Thus, students learned more in the Revoicing condition, and the effect was not specific to a learning objective. There was also no significant effect of condition that remained by the Postdiscussion test, which demonstrates that whatever advantage students in the Revoicing condition achieved during the activity, the other students were able to catch up while interacting with them in the whole class discussion that followed.

The process analysis using the random intercept and slope model showed an interesting contrast between the two interventions that is indicative of a possible explanation for the differential effect on learning during the collaborative activity. With the Revoicing agent, we saw the pattern that we anticipated in conjunction with a positive learning effect. There was no significant difference in intercept between conditions, confirming that there was no difference in absolute number of revoicable assertions between conditions. More importantly, there was no significant correlation between the number of revoicable assertions of a student and that of his partner students in the control condition where there was not a Revoicing agent. However, there was a significant interaction between the condition variable and the number of revoicable assertions contributed by partner students $(\mathrm{R} = 0.14, z = 2.03, p < 0.05)$, indicating that in the Revoicing condition, there was a significantly higher positive correlation between the number of revoicable assertions contributed by a student and that contributed by partner students. Thus, we do see evidence that the intervention had the effect of precipitating pockets of intensive discussion.

In contrast, with the Feedback intervention we see an entirely different pattern. In this case, there was a significant positive effect on the intercept associated with the Feedback condition, indicating that students contributed significantly more revoicable assertions in the Feedback condition; however, there was a marginal interaction between condition and the number of revoicable assertions, this time with a negative coefficient $(\mathrm{R} = -0.16, z = -1.87, p = 0.07)$, indicating that while students were talking more, they were interacting with one another less intensively, which is consistent with the finding of no effect on learning. A possible explanation is that the Feedback agent elicited interaction with itself while the Revoicing agent elicited interaction between students, which was the goal.

4.5 Discussion

The results offer support for the first two hypotheses, namely that one form of APT-based support increases the intensity of interaction between students and increases learning during the collaborative activity. We do not find support for the third hypothesis, that it better prepares students for learning during a whole class discussion. In contrast, what we see is that students who learned less during the collaborative activity caught up with the students who learned more when they all interacted together in the whole class discussion.

Another interesting finding from this study is the differential effect of the two distinct APT manipulations. Whereas Revoicing had a positive effect on learning as well as on the intensity of the interaction, Feedback had no effect on learning and a marginally negative effect on the intensity of the interaction. Further investigation into the nature of the discussions that took place in the different conditions will be needed to understand how the manipulations lead to differing effects.

5 Conclusions and Current Directions

This paper presents a first successful evaluation of a new form of dynamic support for collaborative learning that was inspired by the work in the classroom discourse community on academically productive talk. This form of support was implemented within a recently developed agent-based architecture called Bazaar, which extends earlier work with the Basilica architecture. The proposed dynamic support approach was evaluated in a classroom study involving seven ninth grade biology classes in an urban school district. The study provides evidence that one form of the support, namely, the Revoicing support, resulted in intensification of discussion within the collaborative learning interaction and significantly more learning during the activity.

Acknowledgments

The authors wish to thank Lauren Resnick, Sandra Katz, Catherine Stainton, and Sherice Clarke for their involvement in the district wide teacher training effort that provides the context for this work. This work was supported in part by US National Science Foundation grant SBE 0836012 to the Pittsburgh Science of Learning Center.

References

[1] P. Adey and M. Shayer, “An Exploration of Long-Term Far-Transfer Effects Following an Extended Intervention Program in the High School Science Curriculum,” *Cognition and Instruction*, vol. 11, no. 1, pp. 1-29, 1993.

[2] H. Ai, R. Kumar, D. Nguyen, A. Nagasunder, and C.P. Rosé, Exploring the Effectiveness of Social Capabilities and Goal Alignment in Computer Supported Collaborative Learning," *Proc. 10th Int'l Conf. Intelligent Tutoring Systems,* 2010.

[3] M. Azmitia and R. Montgomery, "Friendship, Transactive Dialogues, and the Development of Scientific Reasoning," *Social Development,* vol. 2, no. 3, pp. 202-221, 1993.

[4] N. Baghaei, A. Mitrovic, and W. Irwin, "Supporting Collaborative Learning and Problem Solving in a Constraint Based CSCL Environment for UML Class Diagrams," *Int'l J. Computer Supported Collaborative Learning,* vol. 2, no. 3, pp. 159-190, 2007.

[5] M. Berkowitz and J. Gibbs, "Measuring the Developmental Features of Moral Discussion," *Merrill-Palmer Quarterly,* vol. 29, pp. 399-410, 1983.

[6] P. Dillenbourg, "Over-Scripting CSCL: The Risks of Blending Collaborative Learning with Instructional Design," *Three Worlds of CSCL: Can We Support CSCL,* pp. 61-91, 2002.

[7] P. Dillenbourg and F. Hong, "The Mechanics of CSCL Macro Scripts," *The Int'l J. Computer-Supported Collaborative Learning,* vol. 3, no. 1, pp. 5-23, 2008.

[8] D. Diziol, E. Walker, N. Rummel, and K.R. Koedinger, "Using Intelligent Tutor Technology to Implement Adaptive Support for Student Collaboration," *Educational Psychology Rev.,* vol. 22, no. 1, pp. 89-102, 2010.

[9] S. Fernando and M. Stevenson, "A Semantic Similarity Approach to Paraphrase Detection," *Proc. 11th Ann. Research Colloquium Computational Linguistics UK (CLUK '08),* 2008.

[10] L. Kobbe, A. Weinberger, P. Dillenbourg, A. Harrer, R. Hämäläinen, P. Häkkinen, and F. Fischer, "Specifying Computer-Supported Collaboration Scripts," *The Int'l J. Computer-Supported Collaborative Learning,* vol. 2, nos. 2/3, pp. 211-224, 2007.

[11] I. Kollar, F. Fischer, and F.W. Hesse, "Collaborative Scripts - A Conceptual Analysis," *Educational Psychology Rev.,* vol. 18, no. 2, pp. 159-185, 2006.

[12] R. Kumar, G. Gweon, M. Joshi, Y. Cui, and C.P. Rosé, "Supporting Students Working Together on Math with Social Dialogue," *Proc. SLaTE Workshop Speech and Language Technology in Education,* 2007.

[13] R. Kumar, C.P. Rosé, Y.C. Wang, M. Joshi, and A. Robinson, "Tutorial Dialogue as Adaptive Collaborative Learning Support," *Proc. Conf. Artificial Intelligence in Education,* 2007.

[14] R. Kumar, H. Ai, J. Beuth, and C.P. Rosé, "Socially-Capable Conversational Tutors Can be Effective in Collaborative Learning Situations," *Proc. 10th Int'l Conf. Intelligent Tutoring Systems,* 2010.

[15] R. Kumar and C.P. Rosé, "Architecture for Building Conversational Agents that Support Collaborative Learning," *IEEE Trans. Learning Technologies,* vol. 4, no. 1, pp. 21-34, Jan. 2011.

[16] P. Lison, "Multi-Policy Dialogue Management," *Proc. Special Interest Group on Discourse and Dialogue Conf. (SIGDIAL '11),* pp. 294-300, 2011.

[17] S. Michaels, C. O'Connor, and L.B. Resnick, "Deliberative Discourse Idealized and Realized: Accountable Talk in the Classroom and in Civic Life," *Studies in Philosophy and Education,* vol. 27, pp. 283-297, 2007.

[18] S. Rabe-Hesketh, A. Skrondal, and A. Pickles, *GLLAMM Manual,* Univ. of California, Berkeley, Division of Biostatistics Working Paper Series, p. 160, 2004.

[19] C.P. Rosé, P. Jordan, M. Ringenberg, S. Siler, K. VanLehn, and A. Weinstein, "Interactive Conceptual Tutoring in Atlas-Andes," *Proc. Conf. AI in Education,* 2001.

[20] C.P. Rosé, Y.C. Wang, Y. Cui, J. Arguello, K. Stegmann, A. Weinberger, F. Fischer, "Analyzing Collaborative Learning Processes Automatically: Exploiting the Advances of Computational Linguistics in Computer-Supported Collaborative Learning," *The Int'l J. Computer-Supported Collaborative Learning* vol. 3, no. 3, pp. 237-271, 2008.

[21] C. Wecker, F. Fischer, "Fading Scripts in Computer-Supported Collaborative Learning: The Role of Distributed Monitoring," *Proc. Eighth Int'l Conf. Computer Supported Collaborative Learning,* pp. 764-772, 2007.

[22] A. Weinberger, K. Stegmann, and F. Fischer, "Scripting Argumentative Knowledge Construction: Effects on Individual and Collaborative Learning," *Proc. Mice, Minds, and Soc.: Int'l Conf. Computer-Supported Collaborative Learning (CSCL '07),* C. Chinn, G. Erkens, and S. Puntambekar, eds., pp. 37-39, 2007.

Gregory Dyke received the PhD degree at the Ecole des Mines de Saint-Etienne in 2009. He then was a postdoctoral reasearcher at Carnegie Mellon University through Fall of 2011. Now, he is a postdoctoral researcher at CNRS in Lyon, France. His doctoral research into frameworks for modeling and capitalizing on analyses of interaction resulted in a student best paper award at CSCL 2009.

David Adamson received the MS degree in language technologies from Carnegie Mellon University in 2012. He taught high school math and computer science in Baltimore, Maryland, from 2004 to 2010. He currently serves as a research programmer at the Language Technologies Institute.

Iris Howley received the bachelor of science degree from Drexel University, where her research centered around applications leveraging the Semantic Web. She is working toward the PhD degree in Carnegie Mellon University's Human-Computer Interaction Institute. Currently, her research interests include the fields of dialogue agents, behavioral psychology, and experimental design.

Carolyn Penstein Rosé received the PhD degree in language and information technologies from Carnegie Mellon University in 1998. She is an associate professor with joint appointments in the Language Technologies Institute and the Human-Computer Interaction Institute at Carnegie Mellon University. She has more than 125 peer reviewed publications, mainly in top tier conferences and journals, spanning the fields of language technologies, learning sciences, and human-computer interaction. She is a member of the IEEE.