Int J Artif Intell Educ (2014) 24:92–124
DOI 10.1007/s40593-013-0012-6

RESEARCH ARTICLE

Towards an Agile Approach to Adapting Dynamic Collaboration Support to Student Needs

David Adamson · Gregory Dyke · Hyeju Jang · Carolyn Penstein Rosé

Published online: 3 January 2014

Abstract This paper investigates the use of conversational agents to scaffold on-line collaborative learning discussions through an approach called Academically Productive Talk (APT). In contrast to past work on dynamic support for collaborative learning, where agents were used to elevate conceptual depth by leading students through directed lines of reasoning (Kumar & Rosé, *IEEE Transactions on Learning Technologies, 4*(1), 2011), this APT-based approach uses generic prompts that encourage students to articulate and elaborate their own lines of reasoning, and to challenge and extend the reasoning of their teammates. This paper integrates findings from a series of studies across content domains (biology, chemistry, engineering design), grade levels (high school, undergraduate), and facilitation strategies. APT based strategies are contrasted with simply offering positive feedback when the students themselves employ APT facilitation moves in their interactions with one another, an intervention we term Positive Feedback for APT engagement. The pattern of results demonstrates that APT based support for collaborative learning can significantly increase learning, but that the effect of specific APT facilitation strategies is context specific. It appears the effectiveness of each strategy depends upon factors such as the difficulty of the material (in terms of being new conceptual material versus review) and the skill level of the learner (urban public high school vs. selective private university). In contrast, Feedback for APT engagement does not positively impact learning. In addition to an analysis based on learning gains, an automated conversation analysis technique is presented that effectively predicts which strategies are successfully operating in specific contexts. Implications for design of more agile forms of dynamic support for collaborative learning are discussed.

D. Adamson · H. Jang
Language Technologies Institute, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA

G. Dyke
CNRS/PRES Université de Lyon, ENS, Lyon, France

C. P. Rosé (✉)
Language Technologies Institute and Human-Computer Interaction Institute, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA
e-mail: cprose@cs.cmu.edu

Keywords Dynamic support for collaborative learning · Automatic collaborative learning process analysis

Introduction

With the recent press given to online education and increasing enrolment in Internet-based courses, the need for scaling up quality educational experiences online has never been so urgent. The biggest limitations are related to the human side of effective educational experiences. While instructor time is a scarce commodity in many such courses, students are plentiful. Thus, one important contribution the field of intelligent support for group learning can make is to develop technologies to structure interactions between students in order to maximize the benefit students receive from one another. Effective collaborative learning experiences are known to provide many benefits to learners in terms of cognitive, metacognitive, and social impact (Kirschner et al. 2009; Scardamalia and Bereiter 1993, 2006; Webb and Palinscar 1996). These experiences offer a potentially valuable resource for massively open online courses, if affordances can be provided that facilitate high quality collaborative learning interactions in the absence of human facilitators that can keep up with the high enrolment in such courses. Effective, automated support for such interactions is the key.

In this paper, we build on a paradigm for dynamic support for group learning that has proven effective for improving interaction and learning in a series of online group learning studies. In particular we refer to using tutorial dialogue agent technology to provide interactive support within a synchronous collaborative chat environment (Kumar et al. 2007; Chaudhuri et al. 2008, 2009; Kumar et al. 2010; Ai et al. 2010; Kumar and Rosé 2011). This form of support can be called dynamic for two reasons. First, the conversational agents are interactive. They have the capability of conducting multi-turn directed lines of reasoning with students that respond to the particulars of student input in response to their prompts (Rosé et al. 2001; Rosé and VanLehn 2005). Second and more importantly, they can be triggered through real time analysis of the collaborative discussion as it unfolds (Kumar et al. 2007, 2010; Adamson et al. 2013; Dyke et al. 2013). The decision making process for identifying triggers in the ongoing collaboration in real time and then launching a specific supportive behaviour at the appropriate time in response to those triggers can be thought of as a strategy. In our prior work, each study described a single strategy that was meant to behave dynamically, according to the same context-sensitive rules for all student groups. In the current work, we explore the ways the dynamic support strategy itself might need to be adapted depending upon the characteristics of the student population. In particular, we build on prior work in triggering support based on real time analysis of collaborative discourse and work towards a new characteristic of dynamic support. Specifically, we are building an empirical foundation for adapting the strategy taken by the support technology to the specific, contextual needs of different student populations. We refer to the concept of strategy adaptation that we work towards in this article as agile support for collaborative learning.

This paper integrates findings from a series of studies across content domains (biology, chemistry, engineering design), grade levels (high school, undergraduate), and facilitation strategies. In each study, each experimental condition makes use of only one strategy. As we observe the pattern of results across studies, where the studies differ in domain and grade level, we see that the ranking among strategies in terms of the relative effectiveness of alternative strategies differs depending on the student population and learning task. We also observe a characteristic pattern in the interaction between students within successful conditions that can be detected with high reliability through automated collaborative process analysis. Thus we offer this series of studies along with the automated process analysis technique as an initial empirical foundation for the development of a more agile approach to dynamic support for group learning. In particular, the choice of facilitation strategy can be adapted in response to an assessment of the patterns of interaction, i.e., whether the characteristic pattern indicating a successful intervention is present.

In the remainder of the paper we first describe a theoretical foundation from prior work in the literature on computer supported collaborative learning, tutorial dialogue agents, and classroom discourse. We then describe our technical approach, which is a publically available architecture for dynamic support for collaborative learning called Bazaar. Next we describe the set of experimental studies we present in this article. Finally, we integrate across the results presented in the individual studies in order to motivate a research agenda for future work in the area of intelligent support for group learning. We conclude with a discussion of the limitations of this study and remaining research questions.

Theoretical Framework

The theoretical foundation for the work reported in this paper comes from three areas. We begin with literature from the Computer Supported Collaborative Learning (CSCL) community. Here we draw insights into types of conversational interactions that are associated with learning in groups and typical static technology for increasing the prevalence of those types of interactions, and thereby increasing learning. Next we review more recent work from the CSCL community where dynamic forms of support for group learning have been developed and demonstrated to be advantageous over more typical static forms of support. We then review the classroom discussion facilitation literature that motivates the set of dynamic support strategies we evaluate in this paper. We propose that these strategies can serve as building blocks for a new form of dynamic support for group learning that we refer to as "agile" support for group learning.

Supporting Effective Collaborative Discussion Using Static Script-Based Support

The field of Computer Supported Collaborative Learning (CSCL) has a rich history extending for nearly two decades, covering a broad spectrum of research related to learning in groups, especially in computer mediated environments. A detailed history is beyond the scope of this article, but interested readers can refer to Stahl's well known history of the field (Stahl et al. 2006) and other foundational work (Dillenbourg et al. 1995).

An important technological goal of work in the field of CSCL is to develop environments with affordances that support effective group learning. The foundation for this work comes from insight into the patterns of conversational interactions that are valuable for learning. A series of studies in the computer-supported collaborative learning field demonstrate the pedagogical value of social interaction from a cognitive perspective, showing that interventions that intensify argumentative knowledge construction, in support of group knowledge integration and consensus building, enhances the development of multi-perspective knowledge (Weinberger et al. 2007; Weinberger and Fischer 2006).

Despite differences in orientation between alternative subcommunities of the learning sciences, some conversational behaviors that have been identified as valuable are very similar across subcommunities. Some such example frameworks for characterizing valuable conversational behaviors share two aspects: namely, the requirement for reasoning to be explicitly displayed in some form, and the preference for connections to be made between the perspective of one student and that of another. It is related to this characterization of valuable discussion behaviors for learning that we base our work in this article. Alternative frameworks for analysis of group knowledge building that privilege subtly different formulations of these behaviors are plentiful. In particular, these include Transactivity (Berkowitz and Gibbs 1983; Teasley 1997; Weinberger and Fischer 2006), Inter-subjective Meaning Making (Suthers 2006), and Productive Agency (Schwartz 1998). Schwartz and colleagues arguing from a Sociocultural perspecive (Schwartz 1998) and de Lisi and Golbeck arguing from a Piagetian Cognitivist perspective (de Lisi and Golbeck 1999) make very similar arguments for the significance of these kinds of behaviors. The idea of transactivity comes originally from a Piagetian framework. It is important to note that when Schwartz describes, from a Vygotskian perspective, the mental scaffolding that collaborating peers offer one another, he describes it in terms of one student using words that serve as a starting place for the other student's reasoning and knowledge construction. This implies explicit articulations of reasoning, so that the reasoning can be known by the partner and then built upon by that partner. The process is explained similarly to how we describe the production of transactive contributions. In both cases, mental models are articulated, shared, mutually examined, and possibly integrated.

The most popular formalization of the construct of transactivity (Berkowitz and Gibbs 1979) includes 18 types of transactive moves. These characterize each student's conversational turn, as long as it is considered an explicit reasoning display that connects with some previously articulated reasoning display. Within this schema, transacts have been divided along multiple different dimensions, which we will draw from later in the article to motivate our series of experimental studies. One important dimension represents whether the transact might be self-oriented (the contribution operates on the speaker's own reasoning) or other-oriented (the contribution operates on the reasoning of a partner) (Teasley 1997; Berkowitz and Gibbs 1979). Another important dimension is whether the contribution represents the original idea as stated or transforms it. Another dimension is whether the contribution is consensus oriented or conflict oriented.

In order to support the growth of student discussion skills, it is necessary to design environments with affordances that encourage transactive behaviors and other valuable learning behaviors. The most popular approach to providing such affordances in the past decade has been that of script-based collaboration (Dillenbourg 2002; Kollar et al. 2006; Kobbe et al. 2007). A script is a schema for offering scaffolding for collaboration. Some typical forms of scripts come in the form of instructions that structure a collaborative task into phases, or structured interfaces that reify certain types of contributions to the collaboration. Such scripts are typically implemented statically, providing the same support in all cases. A 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 patterns of interaction between the participants. Static scripts do not behave differently depending on what is happening in the collaboration per se. Instead, they operate according to choices that are made ahead of time and generally held constant within conditions in an experimental study.

Scripts can be classified as either macro-scripts or micro-scripts (Dillenbourg and Hong 2008). Macro-scripts are pedagogical models that describe coarse-grained features of a collaborative setting, which sequence and structure each phase of a group's activities to foster learning and social interaction. Micro-scripts, 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. Scripts can be more or less coercive, from strict "follow-me" style prompts to subtle suggestions of behavior implicit in the activity's structure. Stricter scripts can work to reduce the gap between expected and observed student behavior, producing a more uniform appearance of discussion. However, they run the risk of over-scripting (Dillenbourg 2002), where the application of inappropriate or unneeded supports have a detrimental effect on collaboration and learning.

Dynamic Script-Based Support with Conversational Agents

The early non-adaptive scripting approaches described above can sometimes result in both over-scripting and in interference between multiple scripts (Weinberger et al. 2007), both of which have been shown to be detrimental to student performance. More dynamic approaches can trigger scripted support in response to the automatic analysis of participant activity (Soller and Lesgold 2000; Erkens and Janssen 2008; Rosé et al. 2008; McLaren et al. 2007; Mu et al. 2012). This sort of analysis can occur at a macro-level, following the state of the activity as a whole, or it can be based on the micro-level classification of individual user contributions. Some prior work on adaptive support for collaborative learning used hint-based support for individual learning with technology to support peer tutoring interactions (Diziol et al. 2010). Other prior work on dynamic conversational agent based support built on a long history of work using tutorial dialogue agents to support individual learning with technology (Wiemer-Hastings et al. 1998; Rosé et al. 2001; Graesser et al. 2002; Zinn et al. 2002).

The collaborative tutoring agents described by Kumar and colleagues (Kumar and Rosé 2011; Kumar et al. 2007) were among the first to implement dynamic scripting in a CSCL environment. In that work, the role of the support was to increase the conceptual depth of discussions by occasionally engaging students in directed lines of reasoning called Knowledge Construction Dialogues (KCDs) (Rosé and VanLehn 2005) that lead students step by step to construct their understanding of a concept and how it applies to the collaborative problem solving context. These encounters were triggered in the midst of collaborative discussions by detection that students were discussing an issue that is associated with one of the pre-authored interactive directed lines of reasoning. Thus, these interventions had the ability to be administered when appropriate given the discussion, rather than being triggered in a one-size-fits-all fashion. In an initial evaluation (Kumar et al. 2007), this form of dynamic support was associated with higher learning gains than a control condition where students had access to the same lines of reasoning, but in a static form. In a subsequent study, students were found to gain significantly more if they had the option to choose whether or not to participate in the directed line of reasoning when it was triggered (Chaudhuri et al. 2009). Scripting such as this offers the potential for minimal interventions to be used more precisely and to greater effect, with greater likelihood of students internalizing the support's intended interaction patterns. Further, the benefits of fading support over time (Wecker and Fischer 2007) could be more fully realized, as the frequency of intervention could be tuned to the students' demonstrated competence.

A major limitation of the specific form of interactive support provided by KCDs is that by their very nature they are content specific. Thus, for every new concept, a separate authoring effort was necessary, which limits the scalability of the approach.

Towards a New Generation of Dynamic Support for Collaborative Learning Inspired by Academically Productive Talk

A promising direction for addressing the issue raised above related to content specificity is to draw inspiration from the classroom discourse literature, where content independent strategies for eliciting valuable interaction between students have been developed and tested. One notable framework for such elicitation is Academically Productive Talk (APT) (Michaels et al. 2008). APT is a classroom discussion facilitation approach that has grown out of instructional theories that emphasize the importance of social interaction in the development of mental processes, in particular ones that value engaging students in transactive exchanges. Drawing on over 15 years of observation and study, Michaels, O'Connor and Resnick propose a number of core "moves" displayed in Table 1. These serve as tools that teachers can employ in order to encourage the development of academically productive classroom discussions – in other words, classroom discussions in which students make their reasoning public, listen deeply and critically to one another's contributions, and then interact with them transactively.

Our recent pilot efforts have begun to develop intelligent conversational agent facilitators whose behavior is not content specific, but rather draws from this literature on facilitation strategies (Adamson et al. 2013; Clarke et al. 2013; Dyke et al. 2013). The design of such support is consistent with the literature on facilitation of collaborative learning groups (e.g., Hmelo-Silver and Barrows 2006), and leverages the large body of work that has shown that APT facilitation behaviors are beneficial for learning with understanding (Adey and Shayer 1993; Bill et al. 1992; Chapin and O'Connor 2004; Resnick et al. 1993, 2013; Topping and Trickey 2007; Wegerif et al. 1999).

The set of Academically Productive Talk moves includes the revoice of a student statement: "So let me see if I've got your thinking right. You're saying XXX?", which encourages students to reformulate or transform the articulation of their reasoning in

Table 1 Academically productive talk facilitation moves

Example teacher utterance	Accountable talk move	Transact category
Explain your thinking.	SAY MORE	SELF ORIENTED, REPRESENTATIONAL, CONSENSUS ORIENTED
What's it prove? Put it into words.	PRESS FOR REASONING	SELF ORIENTED, REPRESENTATIONAL, CONSENSUS ORIENTED
Let me see if I understand correctly. Are you saying they were all adopted?	REVOICE	SELF ORIENTED, TRANSFORMATIONAL, CONSENSUS ORIENTED
If capital 'G''s dominant, wouldn't all babies be orange?	CHALLENGE	SELF ORIENTED, TRANSFORMATIONAL, CONFLICT ORIENTED
Can you repeat what she said?	RESTATE	OTHER ORIENTED, REPRESENTATIONAL, CONSENSUS ORIENTED
Help him out, Stephen. Can you add to what he said?	ADD MORE	OTHER ORIENTED, REPRESENTATIONAL, CONSENSUS ORIENTED
Kelly, are they right? Do you agree or disagree with what they said?	AGREE/ DISAGREE	OTHER ORIENTED, REPRESENTATIONAL, CONFLICT ORIENTED
In your own words, explain why she's right or wrong.	EXPLAIN OTHER	OTHER ORIENTED, TRANSFORMATIONAL, CONFLICT ORIENTED

order to clarify their meaning. Another move involves asking students to apply their own reasoning to someone else's reasoning: "Do you agree or disagree, and why?", which may stimulate sociocognitive conflict, otherwise known as conflict-oriented consensus building. As we have illustrated in Table 1, these core moves can be characterized in terms of the type of transactive behavior they might elicit from students along the three dimensions we introduced above. It is important to note that across these dimensions, these types of transacts can be seen as having a logical ordering which might then apply to the corresponding APT facilitation moves as well. For example, one must understand one's own reasoning before one can hope to understand another person's reasoning, thus self-oriented transacts could be seen as less demanding than other-oriented ones. Furthermore, one must understand reasoning as stated before one can transform or extend that reasoning, thus representational transacts might be seen as less demanding than transformational ones. Reasoning must be understood before it can be rightly challenged, thus, it would be possible to argue that conflict oriented consensus building requires more than consensus oriented transactive behavior. Some prior work has attempted to tease apart differential meditational effects of transacts from these various categories (Azmitia and Montgomery 1993). Building upon this foundation, it is reasonable to hypothesize that the specific APT move that would be helpful to students would depend upon the student's specific capabilities or the difficulty of the material being discussed.

In earlier published 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 (Adey and Shayer 1993; Bill et al. 1992; Chapin and O'Connor 2004; Resnick et al. 1993, 2013; Topping and Trickey 2007; Wegerif et al. 1999). These successes in the classroom discourse literature offer hope that these facilitation strategies could be used to design effective support for collaborative learning, a concept we refer to as APT agents. However, none of these earlier studies have explored the question of what the preconditions for successful use of specific APT moves might be, or what kinds of learners would benefit most from which facilitation moves. Nevertheless, this kind of detailed insight is needed if these moves are to be used to their maximum benefit as support for collaborative learning.

We report on the first wave of APT agent studies in this article. Rather than treat the conversational agents as the sole participants with enough authority to direct the discussion, we encouraged the students to practice productive talk themselves. Thus in each study we offered the students instruction on APT moves prior to their online interaction with one another. The agents' use of APT moves could then serve both to model this style of discussion, as well as to directly facilitate transactive conversational behaviour between students. We also included an intervention to offer encouraging feedback to students for either using APT moves, or engaging in the behaviors the moves were meant to elicit.

In an initial published proof of concept regarding the effectiveness of APT agents at improving collaborative processes and learning (Dyke et al. 2013), the collaborative task was to engage in a series of collaborative discussions in which students make predictions, then make observations, and then explain why their predictions did or did not come to pass. Through this experience, the 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. The collaborative task content, the macro-scripts that supported it, and the list of key concepts used for revoicing were all developed iteratively with feedback from teachers and content experts. An excerpt from the study that shows the agent offering an APT move in its first turn and a Feedback move in its second turn is displayed in Table 2. The tutor's

Table 2 Example of feedback and revoicing in the Dyke et al. study

Time	Author	Text
08:41	Student1	it weighs more because there is more in it
08:45	Student2	starch is tested with a tube as the glucose is tested with a piece of paper like material
08:45	**Tutor**	**Thanks for offering an explanation, Student1 :-)**
08:46	Student3	the longer you leave the test strip in the water the darker green the strip gets and the more weight the glucose solution collects
09:22	Student1	Student3, wouldn't it just show that there was more in it
09:26	**Tutor**	**Would another way to say that be "indicators can prove that there was a change in concentration ?"**
09:31	Student2	Wat Student3 said and starch cant get any darker when purple and the water would be clear so no more

feedback move is triggered by Student1's explanation attempt in the first turn. The tutor's revoicing move is triggered by Student3's contribution in the fourth turn.

In this study, the APT agent provided both macro scripting and micro scripting support in order to structure the interaction. The macro scripting support provided a common task structure across conditions. While acting in the role of macro support provider, the APT agent provided instructions for the collaborative task, and introduced each step of the collaborative task, with the goal of controlling for time on task across conditions. This behavior is not displayed in the excerpt above. The micro level support was meant to respond to the particulars of the conversation as it unfolded. Each experimental condition was defined based on which behaviors would trigger a supportive move, and what that move would be.

The Dyke et al. (2013) study was run as a 2 × 2 between subjects factorial design in which the interactive support provided some behaviors in common across conditions, but other behaviors were manipulated experimentally. The first variable for manipulation was the presence or absence of the Revoicing behavior. The second variable was the presence or absence of the APT Feedback behavior, which is simply positive reinforcement when students were detected to engage in APT behaviour with one another. Students showed significant learning gains in all conditions, and there was a significant main effect of Revoicing such that students in the Revoicing condition learned significantly more between Pretest and Posttest, with an effect size of 0.34 standard deviations. There was no significant main effect of Feedback although there was a trend for it to have a negative effect. And there was no significant interaction between the two factors.

Despite the substantial literature supporting the effectiveness of APT in classroom discussions, it must be acknowledged that much is not known about the mechanism through which the complex intervention has done its work. This can only be determined through more fine-grained, careful experimentation. The treatment has always been complex involving multiple facilitation moves, used within whole classes, where a human teacher insightfully decides when and with whom to use each move. The series of controlled studies presented in this article was meant to begin to fill this empirical gap, in order to begin to build an empirical foundation for evidence-based design principles for development of effective APT-inspired dynamic support for collaborative learning in groups. The Dyke et al. study is the first study that demonstrated the effectiveness of Revoicing as support for collaborative learning with 9th graders, and thus it forms the starting place for our series of studies investigating the generality of the effect in this article.

Bazaar: A Flexible Architecture for Collaboration Support

The publically available Bazaar architecture[1] enables easy integration of a wide variety of discussion facilitation behaviors that has enabled the set of experimental studies we describe in the next section. We begin this section by describing from a user perspective one integrated environment where Bazaar provides collaboration support to distributed groups of learners collaborating synchronously. Next we

[1] http://www.cs.cmu.edu/~dadamson/bazaar/

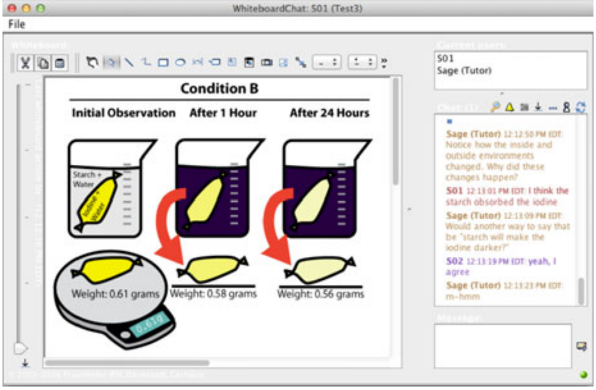
describe the inner workings of the architecture and how it enables effective coordination of supportive facilitation behaviors. We then discuss how we have used this resource to implement the facilitation behaviors we evaluate in our experimental studies.

APT Inspired Dynamic Collaborative Learning Support

The Bazaar architecture (Adamson and Rosé 2012) has been used in a variety of studies (Howley et al. 2012; Clarke et al. 2013; Adamson et al. 2013; Dyke et al. 2013) to implement supportive interventions involving conversational chat agents that participate as facilitators in collaborative learning tasks. The architecture has been successfully integrated with a variety of collaborative environments. These include a standard interface for XMPP multiparty chat, a specialized text chat room with a shared whiteboard (Mühlpfordt and Wessner 2005; Hohenwarter and Preiner 2007). Figure 1 displays an integration between Bazaar and the ConcertChat (Mühlpfordt and Wessner 2005) synchronous chat collaboration environment, which was used in the Dyke et al. (2013) study. Because the Bazaar architecture enables quick development of supportive interventions, one can efficiently proceed from a concept for a new support behavior to a fully functional collaboration environment. In Fig. 1, the panel on the right hand side of the interface is a chat panel where students interact with one another through synchronous chat. The turns labelled as "Tutor" are turns that come from the intelligent conversational agent providing facilitation moves in the conversation. In this example we see the agent performing a Revoicing move. On the left is a shared white board where either the agent or the students can insert images that are then visible to the whole group. In this case, the image displays a cell model that the students were meant to discuss in the Diffusion Lab. The relative size of the chat panel and the white board can be adjusted by clicking in between the two panels and dragging in one direction or the other.

Bazaar

Bazaar is a modular framework for designing multi-party collaborative agents that builds upon the earlier Basilica architecture (Kumar and Rosé 2011). Like Basilica, in



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 absorbed 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

Sage (Tutor) 12:13:23 PM EDT: m-hmm

Fig. 1 CSCL environment from the Dyke et al. study

addition to its core architecture, Bazaar plays host to a library of reusable behavioral components that each trigger a simple form of support. More complex supportive interventions are constructed by integrating multiple simpler behaviors. For example, in the Dyke et al. (2013) study, in the condition with both Revoicing and Feedback, the agent needed to coordinate the macro-level prompts with the micro-level prompts from both the Revoicing and Feedback strategies.

Both the agent's overall composition and the configuration of each component are specified in plain-text properties files, offering a glimpse at the sort of low-overhead flexibility for authoring, content, and deployment championed by recent work (Kobbe et al. 2007). Bazaar and its predecessor are event-driven systems in which independent behavioral components receive, filter, and respond to user, environment, and system-generated events, and present the unified output of these components to the user. Bazaar improves on the Basilica architecture by integrating the orchestration of otherwise competing or conflicting agent behaviors, by simplifying the relationships between components, and by offering an extensible mechanism selecting proposed agent actions. The issue of potential clash between macro-level support and micro-level support is especially important, as we have observed that experiencing these clashes is distracting and confusing for students (Howley et al. 2013). Thus, it is important to note that coordination between simple support behaviors is necessary even when only one APT facilitation strategy is being used.

Figure 2 illustrates a typical Bazaar configuration where events triggered by student contributions in the chat or whiteboard are aggregated in the Input Coordinator. Unlike Basilica, event processing in Bazaar is divided into two distinct phases.

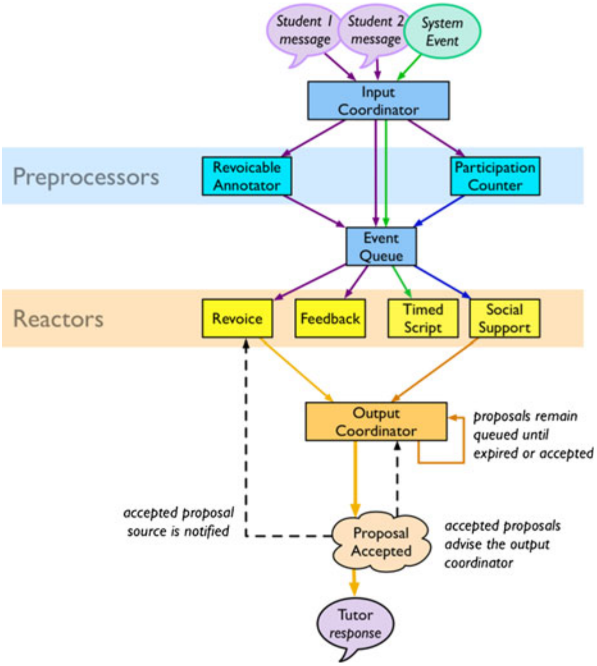


Fig. 2 In the Bazaar pipeline, events are processed in two stages

Preprocessor components analyse the event stream in search of triggers for supportive interventions. Two examples are shown in Fig. 2, including the Revoicable Annotator, which looks for student turns that could be revoiced by the agent, and the Participation Counter, which keeps track of how many utterances each student has contributed recently. These preprocessed events are relayed to a set of Reactors components. Depending on the active agent strategies, under specified circumstances, these Reactors will propose tutor actions in response to these events. The Output Coordinator, described in the next section, then makes decisions about sequencing and timing and thus manages the coordination of potentially clashing interventions. Thus, the Output Coordinator controls when the prompts or other behaviors associated with a triggered strategy are presented to the students.

The Output Coordinator houses Bazaar's primary architectural improvement. In an agent able to offer multiple dynamic behaviors, more than one support strategy may be simultaneously appropriate. Bazaar's predecessors sometimes suffered from clashes between behaviors in cases where multiple were triggered simultaneously. It is important to note that the interference of multiple supports caused by these clashes could invalidate the benefit of any of them, to the detriment of the learner (Weinberger et al. 2007; Howley et al. 2013). It is important to note that participants in a collaborative session, including the facilitator, are not simply focused on the task—they are involved in numerous simultaneous processes including social bonding, idea formation, argumentation, time management, and off-task activity. Managing an APT discussion poses additional challenges. While the kind of in-depth discussion that APT elicits is valuable for learning, it takes time. Facilitators must always keep time constraints in mind in order to achieve an appropriate balance of breadth and depth within and across topics as well as in parcelling out attention to different students.

As we have alluded to, we observed problems with time management in an earlier prototype implementation of an APT agent implemented using Basilica (Howley et al. 2013) that manifested as clashes between the macro and micro scripting behaviors triggered during the study. As a technical solution to this multi-policy management problem, Bazaar draws on and extends the "concurrent mode" approach described by (Lison 2011). In Lison's work, the author adds a "soft" constraint on new proposals by increasing the relative weight of those from the same source as recent actions, preferring that source as a "focus of attention" for as long as it had new actions to propose. Proposals with a great enough activation weight (or priority) from different sources can outweigh this preference, allowing flexible yet consistent responses in the face of noisy input or multiple valid states. Evaluation in a simulated human-robot learning task showed that this "soft" control method performed better than using a hierarchical finite-state controller to select the next source of action. We apply this approach in Bazaar, allowing recent actions to influence the priority of new proposals, and extend it, allowing recent actions to promote or suppress proposals from any source.

In the subsections that follow, we describe Bazaar's event flow in more detail, and the way in which it affords flexible orchestration between multiple behavioral components. This orchestration is key to providing agile, responsive conversational supports. It also underpins Bazaar's role as a rapid research platform.

Events and Components

In Bazaar, an *Event* is an object representing something interesting that has happened in the world of the agent. Some Events come from the environment and map to the actions of participants, like a user entering a chat room, or an incoming user message—these may be annotated by Preprocessor components to reflect a rich understanding of the Event. New Events can also result from the analysis of other Events, or represent awareness of system state. Events such as these are used to launch phases of macro-scripts, or to initiate dynamic support. Bazaar components can generate and respond to arbitrary author-defined Events, thus it is not possible to provide a comprehensive list. The default Event classes handled by the core Bazaar components include Message (a chat message is sent by a student), Presence (a student enters or leaves the chat room), Whiteboard (a student manipulates an object in the shared whiteboard), Dormancy (a student or group has been idle for a certain amount of time), Launch (author-specified conditions for beginning a macro-script have been met), and Step Done (a stage in a macro-script step has been completed).

Components in Bazaar represent a modular representation of related behavior and state-knowledge, corresponding to all or part of a single method of scripting or support. Components respond to those Events they consider relevant. Bazaar defines a two-step event-processing flow, dividing components' event-processing responsibility into *Preprocessor* and *Reactor* roles. While some components may act in both roles, this two-stage processing is still enforced. When a new Event is received by the system, all Preprocessor components that have registered for a particular Event class are given the opportunity to respond to it. They may respond by generating new Events (perhaps to indicate a shift in the conversation's focus) or by adding information to the original Event. Events are subsequently delivered to those Reactor components which are registered for these Events' classes. Reactors have the opportunity to respond to preprocessed Events (to dynamically enact sub-scripts or supports) by proposing actions to the *Output Coordinator*.

Output Coordinator: Prioritizing Proposed Actions

As mentioned above, the Output Coordinator is needed to avoid clashes between multiple proposals that may have been triggered within the same period of time. Most commonly, clashes occur between proposals related to macro level support and proposals related to micro level support. Figure 3 illustrates an example proposal flow within the Output Coordinator. *Proposals* for agent action, received from the Reactor components, are queued by Bazaar's *Output Coordinator*. When a Reactor creates a Proposal, it is assigned a timed window of relevance, and a priority (between 0 and 1). Periodically, the Output Coordinator will re-evaluate the priority of each remaining Proposal (by taking hints from recently enacted Proposals), rejecting those that have expired, and accepting and enacting the Event with the highest priority. A previously-accepted agent action can leave a lingering presence with the Output Coordinator, a *Proposal Advisor*, which can re-weight the priority of (or entirely suppress) incoming Proposals until its influence expires. Each action Proposal is constructed with a timeout-window after which it is no longer relevant—if a queued

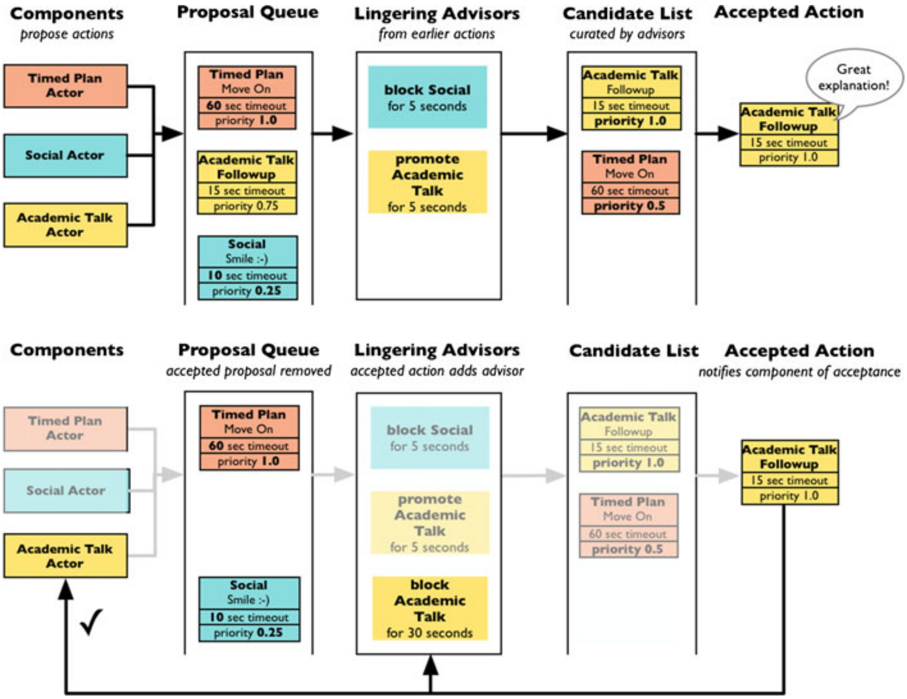

Components
propose actions
Proposal Queue
Lingering Advisors
from earlier actions
Candidate List
curated by advisors
Accepted Action
Timed Plan Actor
Social Actor
Academic Talk Actor
Timed Plan Move On 60 sec timeout priority 1.0
Academic Talk Followup 15 sec timeout priority 0.75
Social Smile :-) 10 sec timeout priority 0.25
block Social for 5 seconds
promote Academic Talk for 5 seconds
Academic Talk Followup 15 sec timeout priority 1.0
Timed Plan Move On 60 sec timeout priority 0.5
Great explanation!
Academic Talk Followup 15 sec timeout priority 1.0
Components
Proposal Queue
accepted proposal removed
Lingering Advisors
accepted action adds advisor
Candidate List
Accepted Action
notifies component of acceptance
Timed Plan Actor
Social Actor
Academic Talk Actor
✓
Timed Plan Move On 60 sec timeout priority 1.0
Social Smile :-) 10 sec timeout priority 0.25
block Social for 5 seconds
promote Academic Talk for 5 seconds
block Academic Talk for 30 seconds
Academic Talk Followup 15 sec timeout priority 1.0
Timed Plan Move On 60 sec timeout priority 0.5
Academic Talk Followup 15 sec timeout priority 1.0

Fig. 3 Proposals are managed by the Output Coordinator

Proposal has not been accepted when its timeout expires, it is removed from the queue. When a message is accepted or rejected, a callback method (which may be defined at the time of Proposal creation) is invoked, allowing the proposing Component to update its state accordingly.

Bazaar provides methods for creating Proposals with Proposal Advisors for common use cases. These include sending simple single turn messages, or interventions that involve sequences of messages and that suppress all subsequent Proposals (or those from a particular set of source components) for a given amount of time or until the sequence of associated behaviors is complete (to allow an opportunity for student follow-up, for example). In most cases, employing these pre-defined advisors is sufficient to author a smooth and natural agent experience. Bazaar also supports more advanced proposal-management techniques, such as affording a Proposal the ability to re-evaluate its own importance in light of subsequent Events.

By allowing Proposals to establish constraints on near-future Events in a general way, conversational agents authored in Bazaar can be responsive to changes in both student behavior, and in the behaviors enacted by the agents' behavioral components. As support behaviors re-evaluate their own relevance, the agent thus has the potential to effectively change strategies dynamically, based on whether the current strategy is having the desired effect. Authors of Bazaar agents can specify these to suit their experimental, pedagogical, and practical needs. In particular, the rigidity of timing with which macro-scripted elements are executed can be adjusted along the spectrum between replicability and internal experimental validity, and natural, external conversational validity. Table 3 details the Proposal and Advisor configurations for components used in the studies described in this article.

Using Bazaar to Implement Supportive Interventions

Three different interventions are evaluated in the series of studies reported in this paper. *Revoicing* elicits Self-Oriented, Transformational, Consensus Oriented transacts. *Agree-Disagre*e elicits Other-Oriented, Representational, Conflict-Oriented transacts. Finally, *APT Feedback* is designed to offer non-specific encouragement for students to engage in APT related behaviors. As will become clear, all of these interventions reused many of the same components in their implementation.

Detecting Academically Productive Talk Candidates

The two APT interventions implemented for the studies reported in this paper required the detection of task-relevant conceptual assertions. For example, attempts at articulation of task-relevant assertions could be the focus of a reformulation elicited by a Revoice facilitation move or the idea that a student agrees or disagrees with in response to an Agree/Disagree move.

In order to identify task-relevant conceptual assertions, we worked with domain experts and instructors to develop a "gold standard" list of statements that captured important concepts and misconceptions for the unit of study. Such statements were drawn from both the experts' knowledge and expectations and from transcripts of an unsupported dry-run of the task. We use a "bag of synonyms" cosine similarity

Table 3 Proposal and Advisor configurations for components described in this article

Bazaar component	Behavior intent	Proposal priority	Proposal timeout	Advisor implementation
Timed script	Provide consistent time for each section across groups, allow time for reading	High	60 s	Block all tutor actions for a time proportional to the length of the displayed prompt.
Social support	Offer immediate responses to social cues	Low	3 s	Block all tutor actions for 5 s.
APT feed-back	Give immediate feedback on student APT behaviors	High	3 s	Block all other tutor actions for 5 s, block other APT moves for 20 s
Revoicing	Highlight and clarify student-generated concepts	Medium *(proportional to candidate similarity)*	15 s	Block all other tutor actions for 10 s, block other APT moves for a further 45 s
Agree disagree	Support discussion of student-generated concepts	Medium *(proportional to candidate similarity)*	15 s	Check for student followup before acting. Prioritize agree-disagree tutor followup prompts. Block other tutor actions for 10 s, block other APT moves for a further 45 s.

measure (Fernando and Stevenson 2008; Mihalcea et al. 2006), which essentially measures overlap in word usage. Student assertions which are within a certain threshold of similarity to the gold statements are identified as *revoicable* or *agree-disagree candidates* that could be evaluated by the group. Both the *Revoicing* and *Agree-Disagree* supports described employ use the same detection method (implemented as a Bazaar Pre-Processor component), although with a looser similarity threshold in the latter case.

Revoicing Facilitation

One of the forms of support evaluated in this paper is a Bazaar agent that performs the APT Revoicing move. The agent compares student input against a list of correct statements drawn from the data collected in pilot runs of the studies. If an entry in this list could be interpreted as a paraphrase of the student's input using the method described above, it is offered by the agent as a "revoicing" to the students. The same statement was never offered more than once in the same session as a revoicing. When student statements were not close enough to match the revoicing list but contained the first mention of important lesson concepts (like "test strip" or "molecule size"), the agent would ask the student or a peer to expand or restate their contribution. Examples are given in Table 4.

An example from a unit of 9th grade biology on Genetics, which was the context for Study 2 discussed below, is displayed in Table 5. Here all of the student turns that are detected to be *revoicable* are marked with italics. The Tutor's revoicing is marked in bold. Note that while two turns were detected as *revoicable* in the Preprocessor, a revoicing was only triggered once because of the constraint that the same concept won't be revoiced more than once in the same conversation. What we see in this example is that the tutor's revoicing of Student1 created the opportunity for that idea to be the focus of reformulation and clarification, as shown by Student2's followup.

Agree-Disagree Facilitation

We also present a conversational agent behavior based on the "Agree-Disagree" APT move. As the group discusses flows, the agent monitors the chat for student assertions that could be followed up by a check for agreement or understanding. This uses the detection method described above, but with a looser match threshold than that

Table 4 Examples of Revoicing behaviors

Student contribution	Revoicing facilitation response
Basically the glucose will get inside	Maybe you could state that as "the cell membrane is permeable to glucose."
It changed because the tube absorbed the iodine,	So are you saying "the molecules diffused through the membrane?"
I predict that if the holes in the plastic are large enough , the glucose will go into the water solution..	Maybe you could state that as "both water and glucose molecules are able to move between the two environments."

Table 5 Extended Revoicing example in a 9th grade biology lesson from a Genetics unit

Time	Author	Text
00:12	*Student1*	*Yes both of the parents are momozygous but yellow cat is dominent and white cat is recessive*
00:27	**Tutor**	**Let me make sure I understand you - are you saying a white cat had to come from parents who both carry the recessive white fur gene?**
00:36	*Student2*	*Because the orange color coat is more dominant than the white color coat*
00:57	Student1	Its true

employed by the Revoicing behavior. This results in the detection of a greater number of candidate statements, and more opportunities for support than the Revoicing agent could afford. Not all of these detected statements will result in the agent triggering an Agree-Disagree move. Instead, when a candidate statement is identified, the Agree-Disagree component waits for the other students in the group to respond to it. If another student responds with an evaluation of their peer's contribution (along the lines of "I agree" or "I think you're wrong"), but does not support their evaluation with an explanation, the agent will encourage this second student to provide one. If a student instead follows up with another candidate statement, the agent does nothing, leaving the floor open for productive student discussion to continue unimpeded, reducing the risk of over-scripting their collaboration. If the other students do not respond with either an evaluation or a contentful followup, the agent prompts them to comment on the candidate statement – for example, "What do you think about Student's idea? Do you agree or disagree?" This interactive process is illustrated in Fig. 4.

Table 6 shows an example of this support in the high-school biology setting. Times are given in seconds from the beginning of the excerpt. Rows marked in italics are the automatically detected labels the agent uses to motivate its facilitation moves. Student1 offers a contribution that is a candidate for evaluation. After 15 s of no students following up, the agent inserts an Agree-Disagree prompt. Student2 then

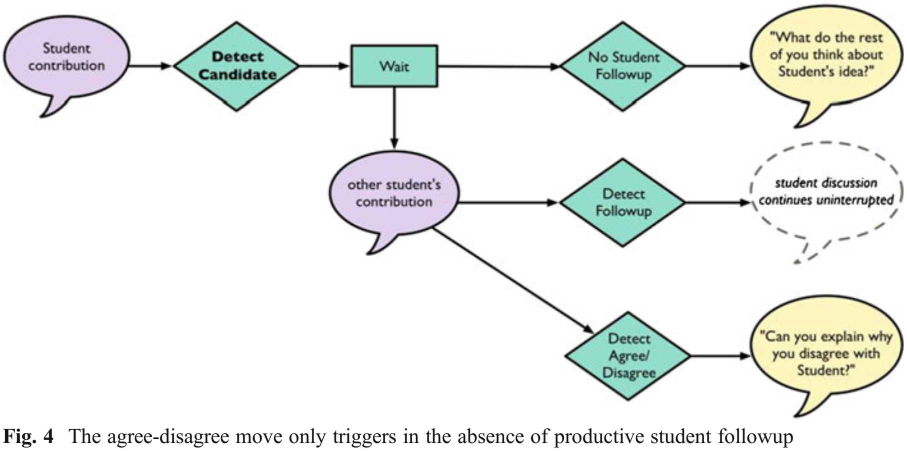

Student contribution
Detect Candidate
Wait
No Student Followup
"What do the rest of you think about Student's idea?"
other student's contribution
Detect Followup
student discussion continues uninterrupted
Detect Agree/ Disagree
"Can you explain why you disagree with Student?"

Fig. 4 The agree-disagree move only triggers in the absence of productive student followup

Table 6 Agree-disagree example in a 9th grade biology lesson from a Genetics unit

Time	Author	Text
00:00	*Student1*	*The yellow cat is probably GG and the white is gg*
00:15	**Tutor**	**Do you agree with Student1? Why, or why not?**
00:17	Student2	Gg,Gg.Gg.Gg.
00:31	Student3	I do shes right
00:47	Student1	I agree with Student2 for the possible outcomes
00:48	Student2	I agree with Student1

responds very quickly, and then Student3 follows suit. Thus, the agent's prompt seems to have successfully engaged the group in responding to Student1.

The excerpt shown in Table 7 is drawn from a unit of college Chemistry in intermolecular forces. This exchange is typical of a group interaction in this environment—Student1's contribution at 17 s, although a candidate for evaluation, is not acted upon by the agent because Student3's follow up pre-empts it. After 15 s following Student3's statement without any sort of uptake by the group, the agent prompts the group to agree or disagree with it. Student2 offers a challenge, which leads to an extended back-and-forth between all three students.

APT Feedback

A final behavior implemented using Bazaar and evaluated in this paper provides positive feedback for student APT. Here the goal is to offer encouragement when students themselves begin to act as APT facilitators with one another. Thus, rather than offer APT facilitation moves, the agent's task here was to detect when students were doing it. Student input was matched against a list of hand-crafted patterns indicating APT moves including explanation, challenges, revoicing, and requests

Table 7 Agree-disagree example in an undergraduate Chemistry unit on intermolecular forces

Time	Author	Text
00:00	Student1	Ok lots of things to do…
00:13	Student1	First one
00:17	*Student1*	*Surface area is higher*
00:20	Student2	Arrow up?
00:22	*Student3*	*Ok, boiling pt will go up and vdW will go up for all of them consecutively… right?*
00:37	**Tutor**	**Do you concur with Student3? Why, or why not?**
00:41	Student2	Hmm not necessarily
00:47	*Student1*	*Area goes up for each*
00:50	Student2	Would it?
00:51	Student2	Im not sure
00:56	*Student3*	*Yea for sure area goes up*

for others to provide each of the same. If a student statement matched, the agent publicly praised the student's move, and (when appropriate) encouraged the other students to attend to and interact with that student. Rather than perform APT based facilitation itself, as the Revoicing behavior does, the Feedback behavior was meant to indirectly support the prevalence of APT in the discussions by encouraging students to take this facilitation role. An example is presented in Table 8, where moves detected as worthy of positive feedback are marked in italics, and the agent's feedback is marked in bold.

Method

The line of inquiry investigated in this article was prompted by the hypothesis that by incorporating intelligent agents to model, support, coach, and provide feedback for students using Academically Productive Talk (APT) moves, students will benefit in terms of learning and interaction. Note that we do not hypothesize that all APT moves are interchangeable. Rather, in this work we manipulate the usage of different APT moves in order to understand better their separate and joint effects on measures of learning and interaction. The experiments presented in this paper build on the early success of a form of APT, namely revoicing support, in a study with 9th grade biology students (Dyke et al. 2013). The series of studies presented in this paper serve as a test of the generality of the effect.

As an advance organizer for the series of studies and analyses, what we will see in these studies is that the positive effect of APT facilitation behaviors is context specific. Thus, a more generalizable form of support would need insights into the contextual pre-conditions for the success of these facilitation strategies. The pattern of results across the studies begins to provide an empirical foundation for a more agile, more generalizable form of support that can use APT facilitation behaviors in a more nuanced, population sensitive way. Note that we are not claiming in our presentation of these studies that we already have this agile form of support. Rather, our investigations provide the initial empirical foundation for developing such an approach. In addition to the learning gains analysis for each study, we present an automated process analysis technique that proves surprisingly accurate in identifying which interventions were most successful in each context. An automated measure that provides an indication of the relative success of alternative intervention strategies

Table 8 Feedback example from the Dyke et al. study

Time	Author	Text
08:41	*Student1*	*It weighs more because there is more in it*
08:45	Student2	Starch is tested with a tube as the glucose is tested with a piece of paper like material
08:45	**Tutor**	**Thanks for offering an explanation, Student1 :-)**
08:46	Student3	The longer you leave the test strip in the water the darker green the strip gets and the more weight the glucose solution collects
09:22	Student1	Student3, wouldn't it just show that there was more in it

within contexts can be used to discover new associations between contexts and facilitation strategies in real time. Thus, we will argue that beyond the insights into the individual contexts investigated in this series of studies, the results allow us to make cautious predictions beyond those contexts using the results from the process analysis we present as one of the contributions of this paper.

Experimental Paradigm Common Across Studies

In all four studies discussed in this paper, which includes the foundational Dyke et al. (2013) study and three new ones, the instructional goal is for students to understand principles that explain causal mechanisms at a deep level. To that end, we prompt students for explanation in the context of group discussion with the goal that students will articulate and monitor their own reasoning, evaluate one another's reasoning, and challenge one another. In all cases, students interact with their group members by logging into a chat room assigned to their group in the ConcertChat environment displayed in Fig. 1 above, a discussion environment with a shared whiteboard (Mühlpfordt and Wessner 2005).

Assessment

In all studies presented in this paper, we employ both summative assessments in the form of pre/post domain-knowledge tests, as well as process assessments that measure the interventions' success in eliciting more of the behaviors that mark effective collaborative learning processes. Thus the first analysis we do in all studies is to verify that learning took place between pre and post-test (using an ANOVA) and then to test for differences in learning between conditions (using an ANCOVA).

Beyond the learning gains analyses, we also do a process analysis. The specific interaction goal of APT interventions is to engage students in a more intensive exchange of explanations. More specifically, the desired contributions within these exchanges are what we referred to above as revoicable assertions. By more intensive, we do not mean that students utter more explanations per se, but that the explanations they utter are directed towards building on those of their partner students. The motivation for attempting to achieve this was to raise the level of critical thinking and learning. Thus, in addition to a Pre/Post test measure of learning, a process analysis to verify that the intervention did its job is also important for evaluating our hypothesis. Anecdotally, we have 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.

In order to accomplish this, the chat logs were segmented into intervals such that one observation is extracted per student for each interval. For young learners, we use 5 min as the interval since they type slower and take more time before responding whereas for older, more advanced learners, we use 2 min as the interval. In this way, we keep the average number of contributions per segment comparable between age groups. In each observation, we count 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 occurrences of revoicable assertions of a student with those of the other students in the same group. We used a multi-level model to analyse the results in order to account for non-independence between instances. We expect to see that the correlation is significantly higher in the condition with the intervention when the intervention is effective. We do the analysis separately for each independent factor within each study in order to contrast discourse behaviour between conditions. 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 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) (Rabe-Hesketh et al. 2004) add-on to STATA (Rabe-Hesketh and Skrondal 2012). 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. A positive difference would be indicative of an intensification of the interaction between students. A significant positive difference in intercept between conditions would indicate that the intervention raised the average number of revoicable assertions within time slices.

Recap of Study 1: 9th Grade Diffusion Lab

The first of four studies, which we discussed above (Dyke et al. 2013), was carried out during a module introducing the concepts of selective permeability, diffusion, osmosis and equilibrium. This study took place in an urban high school, and the content was relatively new to the students since they were at the beginning of a new unit in their course. In this study, the students worked together in a collaborative session for about 20 min. As mentioned, this study was run as a 2 × 2 between subjects factorial design, where the first variable for manipulation was the presence or absence of the Revoicing behavior. The second variable was the presence or absence of the APT Feedback behavior. Students showed significant learning gains in all conditions, and there was a significant main effect of Revoicing such that students in the Revoicing condition learned significantly more between Pretest and Posttest, with an effect size of 0.34 standard deviations. There was no significant main effect of the APT Feedback manipulation although there was a trend for it to have a negative effect. And there was no significant interaction between the two factors.

In order to compare the results from this study with those of the other studies, we present now the process analysis results from this study. The process analysis using the random intercept and slope model showed an interesting contrast between the

Revoicing intervention and the APT Feedback intervention that is indicative of a possible explanation for the differential effect on learning during the collaborative activity. In the Revoicing condition (where there was a Revoicing agent to offer micro level support), we saw the pattern that we anticipated in conjunction with a positive learning effect in comparison with the Control condition (where were was no Revoicing agent). 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 Revoicing condition variable and the number of revoicable assertions contributed by partner students ($R=0.14$, $z=2.03$, $p<.05$). This indicates that there was a significantly higher positive correlation between the number of revoicable assertions contributed by a student and that contributed by partner students in the Revoicing condition. Thus we do see evidence that in the Revoicing condition, the intervention had the effect of precipitating pockets of intensive discussion.

In contrast, with the APT Feedback intervention we see an entirely different pattern. In this case, there was a significant positive effect on the intercept associated with the APT Feedback condition, indicating that students contributed significantly more revoicable assertions in the APT Feedback condition. However, there was a marginal interaction between condition and the number of revoicable assertions, this time with a negative coefficient ($R=-0.16$, $z=-1.87$, $p=.07$). This indicates 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 between students and itself while the Revoicing agent elicited interaction between students, which was the goal.

Study 2: 9th Grade Genetics

The second study was conducted within the same course where the first study was conducted, but 2 months later, in a unit on Genetics. The study was carried out during a module specifically introducing the concept heredity, and the use of Punnet squares as a tool to reason about the inheritance of single traits. At the time of the study, the material was somewhat familiar to the students since they were towards the end of the unit by the time the study took place. In the collaborative activity that lasted for about 20 min, student groups were presented with a set of three problems and asked to reason about the physical and genetic traits of the hypothetical parents of a set of sibling organisms. Specifically, in each problem, students were shown a litter of eight kittens that varied in fur color (either orange or white), and were instructed to identify the genotypes and phenotypes of the parents, and to explain their reasoning to their teammates. This sort of "backwards" reasoning had not been explicitly addressed in the course to date—students only had prior experience with "forward" reasoning from given parental traits. The mystery parents were presented as the inputs to an unpopulated Punnet square, as shown in Fig. 5. As an incentive, students were told that the best team, determined by a combination of discussion quality and post-test scores, would be awarded with a modest prize of food. Each of the three tasks was

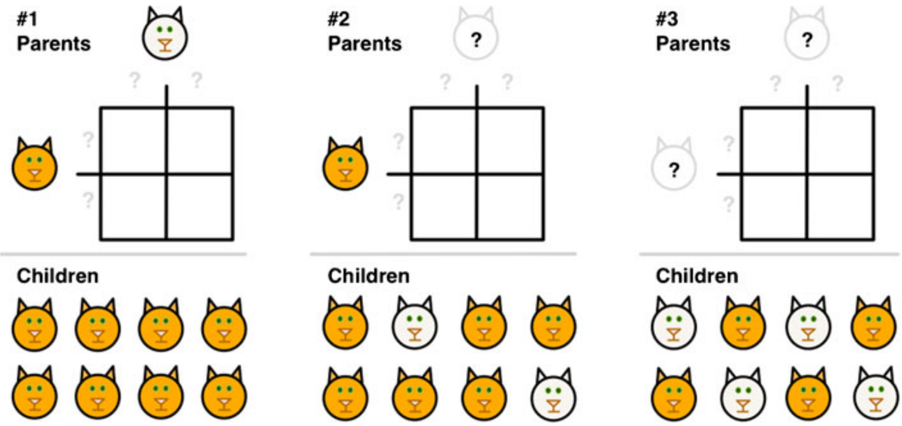


Fig. 5 Example of a Punnet square

progressively harder than the last in that fewer clues about the parent's identities were included.

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

Participants

This study was conducted in the same seven 9th grade biology classes of an urban school district that the first study was run in, only 2 months later. 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.

Experimental Manipulation

In this study, only Revoicing behaviors were manipulated experimentally. The APT Feedback that was evaluated in the first study was not repeated in the second study since it did not lead to a positive effect with this student population in that study. In both conditions of this study, 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. Only the micro-level support varied between conditions.

Study Procedure

Just like in the first study, the students first participated in a normal class lesson on genetics as part of the course curriculum. At the end of the period, they took a pre-test. The pre-test included four multiple-choice questions testing the students' ability to use Punnet squares to reason about the likelihood of genetic and physical traits of children based upon the traits of the parents, and one open-ended question designed to elicit explanation of reasoning about parental identity based upon the physical traits of offspring.

In the next class period, the students participated in a 20 min collaborative computer-mediated activity during which the experimental manipulation took place. The students did the activity in groups of three students, scaffolded by conversational agents. Students within classes were randomly assigned to groups and then groups to conditions. As in the first study, this activity was introduced by a cartoon depicting the use of APT and a reminder of the basic science principles underlying the activity, in this case principles of simple inheritance. At the end of this second phase, the students took a post-test of the same design as the pre-test, although with different characteristics and genotypes presented in each problem.

Results

As in the Diffusion Lab study, we evaluated pre-to-post test learning and the effect of condition on learning and on the collaborative process. However, the material appears to have been too easy for the students. Post-test scores were higher on average than pre-tests scores, but not significantly. And although the trend was for students in the Revoicing condition to learn more than students in the Control condition, the difference was not significant or even marginal. Thus we do not elaborate on the learning gains analysis here.

While the learning gains analysis does not allow us to draw new insights about learning, we can observe how the collaborative processes play out with the same student population used in Study 1, but with material that appears to be less challenging for them. The process analysis using the random intercept and slope model showed an interesting contrast between this study and the Diffusion lab study. Similar to the Diffusion study, there was no significant difference in intercept between conditions, confirming again that there was no difference in absolute number of revoicable assertions between conditions. This time, however, there was a significant correlation between the number of revoicable assertions of a student and that of his partner students in both conditions ($R=0.31$, $z=3.59$, $p<.001$), and no difference in slope between conditions. Thus, we have confirming evidence that there was no difference in effect between conditions. Students were interacting productively in both conditions regardless of support, possibly because the material was easy for them and thus they may not have needed the revoicing support.

Study 3: Freshman Engineering Design

As a second replication of the successful Diffusion Lab study, we ran a study in a Freshman Engineering Design course at a selective private university. The material presented in the study was relatively familiar to the students. The experimental manipulation was identical to that of Study 1, including both the APT Feedback manipulation and the Revoicing manipulation.

Participants

One hundred nine mechanical engineering students participated in the experiment, which was held over six sessions spread evenly between 2 days. Students were grouped into teams of three or four individuals. The number of three person and four person groups was roughly evenly distributed between conditions. In each session, the groups were evenly distributed between the three conditions. The 2 days of the experiment were separated by 2 weeks.

Experimental Procedure

Each session started with a follow-along tutorial of computer-aided analysis where the students analysed a wrench they had designed in a previous lab. A pre-test with 11 questions (7 multiple choice questions and 4 brief explanation questions) was administered after the analysis tutorial. The experimental manipulation happened during the Collaborative Design Competition after the pre-test. Students were asked to work as a team over 90 min to design a better wrench taking three aspects into consideration: ease of use, material cost and safety. Students were instructed to make three new designs and calculate success measures for each of the three aspects under consideration. As part of this process, students occasionally were requested to make predictions and explain them, however, it should be noted that this task was somewhat less conceptually oriented than that used in the other studies.

Results

The results of this study were strikingly different from the two conducted in 9th grade Biology. In particular, rather than achieving a positive effect, the Revoicing manipulation had a significant negative effect on learning within the APT Feedback condition with this more advanced population of learners.

As in the earlier studies, we began our analysis by first verifying that students learned between pre and posttest. For this analysis, we treated Test as a repeated measure, with Pre and Post being the two time points. We conducted an ANOVA test with Test as the dependent variable. Time point, Revoicing, and Feedback were independent variables. We included all two-way interaction terms as well as the three-way interaction term. There was a significant main effect of Time point F(1,210)=9.28, $p<.005$, demonstrating that students learned. None of the interaction terms were significant. Thus students learned between pre and posttest regardless of condition.

Next we tested for differences in learning between conditions. For this analysis, we conducted an ANCOVA with Post-test as the dependent variable and Pre-test as a covariate. Revoicing and APT Feedback were the two independent variables. We also included the interaction term in the model. Here there was almost no effect of APT Feedback F(1, 104)=0.03, $p=.87$. There was a trend for a negative effect of the Revoicing manipulation F(1, 104)=2.22, $p=.13$. The interaction between APT Feedback and Revoicing was not significant, however, it should be noted that within the APT Feedback condition, there was a significant negative effect of Revoicing ($p<.05$). Thus, there is some qualified evidence of a potential detrimental effect of Revoicing with this population.

Consistent with the negative trend, the process analysis using the random intercept and slope model showed an interesting contrast with the earlier studies when we evaluated the effect of the Revoicing manipulation. Similar to the earlier studies, there was no significant difference in intercept between conditions, confirming again that there was no difference in absolute number of revoicable assertions between conditions. There was, however, a significant correlation between the number of revoicable assertions of a student and that of his partner students in the control condition (R=0.1, z=3.7, p<.001), as well as an interaction between condition and slope. In contrast to the Diffusion study where we saw a positive effect of revoicing both on learning and on the slope, here we see a negative impact on slope based on the correlation on the interaction term. This echoes the trend for a negative effect on learning (R=−0.1, z= 2.4, p<.05). Thus, we have confirming evidence that there was a negative impact of the Revoicing manipulation with this population. When we do the same analysis to evaluate the effect of the APT Feedback condition, we see no effect of any variable.

Study 4: Freshman Honors Chemistry

In the final study, published as a conference paper (Adamson et al. 2013), we tested the hypothesis that one reason why Study 3 was not successful was that the students did not need support in making themselves clear. Instead, we hypothesized that instead of support for basic articulation of ideas, they needed support to the next step of challenging each other's reasoning. We consider this study to be a good comparison case to Study 3 because the student population was similarly university level from the same selective private university, and the material was similarly relatively familiar to the students.

The collaborative task, which lasted for about 90 min, focused on intermolecular forces and their influence on the boiling points of liquids. For each problem in the activity, students were asked to predict whether a given substance would have a higher or lower boiling point than two of its relatives, explaining their reasoning about the set of molecules in terms of their structure and the forces at play. Each problem of this sort was followed up by revealing the actual boiling point of the mystery molecule, and asking students to revisit their predictions and explanations in light of the new data. A liquid's boiling point can be influenced simultaneously by a number of different intermolecular forces, each of which arises as a consequence of the molecules' particular structural attributes. Correctly identifying the pertinent structural features of molecules and reasoning about how they will affect the liquid's boiling point is a non-trivial and multi-faceted task. Because multiple types of intermolecular forces influence liquids' boiling points, we employed the Jigsaw technique (Aronson et al. 1978), assigning students within each group to read individually about one of three forces that contribute to a molecule's boiling point. This division also provided intrinsic motivation for collaboration, as the task could not be completed without knowledge from each of the student experts.

Participants

The participants in our study were first-year undergraduate students studying intermolecular forces in an Honors Chemistry course. Students were randomly assigned to groups of three or four, and then groups were randomly assigned to conditions. The balance of three and four person groups was even between conditions, and there was no effect of team size on any of our dependent measures. All students in the course were required to participate in the online exercise for course credit, but they had the option of not consenting for their data to be included in our research. Thus, we only report results for consenting students. Altogether, our analysis includes data from 18 students from 6 different groups, which is 9 students and 3 groups in each condition.

Experimental Manipulation

Our experimental design was a simple 2-condition between-subjects design where teams were assigned randomly either to the Agree-Disagree condition or the Control condition. Both conditions were identical except for inclusion of the Agree-Disagree facilitation move by the agent. Thus, both conditions benefitted both from macro-level and micro-level script based support. In the Agree-Disagree condition, whenever the agent was not engaged in a directed dialog, it was receptive to opportunities to dynamically offer support using the Agree-Disagree behavior, discussed above.

Experimental Procedure

The experimental procedure was simple. Students took a pretest, then participated in pairs in the online collaborative activity, and finally completed a post-test. Pre and post tests were used to measure learning during the collaborative exercise.

Results

Our hypothesis was that the introduction of the Agree/Disagree agent would intensify the interaction between students, which might increase critical thinking, and subsequently increase learning. Our analysis offers qualified support for the hypothesis.

As before, we began our analysis by first verifying that students learned between pre and posttest. For this analysis, we treated Test as a repeated measure, with Pre and Post being the two time points. We conducted an ANOVA test with Test as the dependent variable. Time point and Revoicing were independent variables. We included the interaction between Time point and Condition as well. There was a significant main effect of Time point F(1,31)=7.58, $p<.01$, demonstrating that students learned. The interaction term was not significant. Thus students learned between pre and posttest regardless of condition. As before, to evaluate the effect of condition on learning, we used an ANCOVA with posttest as the dependent variable, pretest as a covariate, Condition as an independent variable. In this analysis, there was a marginal effect of Condition on learning (F(1,11)=1.82, $p<.1$, effect size 0.55 standard deviations), such that students in the Agree/Disagree condition learned more. The effect was moderate.

Next we examined the intensifying effect of the intervention on the interaction between students using the same random intercept and slope model approach used in the earlier studies. The analysis showed the pattern that we expected. 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 an Agree/Disagree agent. There was, however, a significant interaction between the condition variable and the number of revoicable assertions contributed by partner students ($R=0.14$, $z=2.03$, $p<.05$). This suggests that there was a significant positive correlation between the number of revoicable assertions contributed by a student and that contributed by partner students in the Agree/Disagree condition. Thus we do see evidence that the intervention had the effect of precipitating pockets of intensive discussion.

Discussion

The pattern of results across studies is consistent with what we expected to see given the connection between types of transactive discussion behavior and how they are related to the three different discussion facilitation behaviors we explored in this paper. In particular, we contrasted Revoicing, which is meant to elicit self-oriented, consensus-oriented transacts, which we have argued should be less demanding and to some extent logically prior to other-oriented, conflict-oriented transacts, which are elicited by Agree-Disagree facilitation moves. It would therefore be consistent to expect that Revoicing moves would be most needed by younger, less sophisticated learners, whereas Agree-Disagree moves would be more appropriate for more advanced learners. In prior studies of the effect of transactivity on learning (Azmitia and Montgomery 1993), the effect was only observed in material that was difficult for learners, thus we would expect that learners who were close to mastery would not benefit substantially from APT. Thus, where material is easy for learners, we would not predict a difference between conditions where we test APT in comparison with other facilitation behaviors or even no facilitation. A summary of results across studies is given in Table 9.

In study 1 where we test Revoicing against Feedback for APT with young learners on material that was difficult for them, we observe a positive effect of Revoicing. In

Table 9 Summary of results across studies

	9th grade diffusion	9th grade genetics	Freshman engineering design	Freshman Honors Chemistry
Experimental manipulation	Revoicing vs no APT, feedback vs no feedback	Revoicing vs no APT	Revoicing vs no APT, feedback vs no feedback	Agree-disagree vs no APT
Learning effect	Positive effect of Revoicing, no effect of feebdack	No significant effect of Revoicing	No main effect but significant negative effect of Revoicing in feedback condition	Marginal positive effect of agree-disagree
Process analysis	Significant positive effect of Revoicing, marginal negative effect of feedback	No effect of Revoicing	Significant negative effect of Revoicing, no effect of feedback	Significant positive effect of agree-disagree

study 2, we test Revoicing again, but this time with material that was easy for the students. Here there was no significant difference between conditions. This contrast is consistent with what we argued above. It is true that since the group of learners was the same in the two studies, the difference in effect could have potentially been related to the fact that the students were already familiar with the support agents. It is clear, however, that re-exposure to the same manipulation does not completely explain the difference in results across these two studies. In the first study, we observed a significant pre to post test gain across all conditions, including the condition where no support was offered beyond the macro level structuring of the activity. In the second study, no significant pre to post test gain was observed in any condition. Rather, both pre and post test scores were high across conditions, which highlights the fact that the material was easy for the students.

Studies 3 and 4 involve more advanced learners on material that was moderately familiar to them. More advanced learners are already good at articulating their own ideas. Thus, Revoicing support is unneeded support for them. Rather, they need to be pushed beyond that to connect to the reasoning of their partner students. We expect then not to see a positive effect in study 3 where we test Revoicing on these advanced learners, and we do expect to see a positive result with Agree-Disagree, which we test in the final study. And we do see this.

The pattern of results with learning gains is as expected from prior work. What is more striking is the picture that emerges when we compare the pattern of results from the learning gains analysis with that from the process analysis. What we see from the series of studies presented in this paper is that the effect of condition on learning gains and on collaborative process provide largely converging evidence across studies. This convergence highlights the value of the simple form of process analysis presented in this article for evaluating in process effect of collaboration support. It shows that this process analysis can be used to gauge whether an intervention is working appropriately with a group of learners. If the process analysis indicates that the strategy is not a good match for the learners, the strategy can be adjusted. The new strategy can then be evaluated in process the same way, and further adjustments can be made. Thus, this simple automated process analysis technique could form the foundation for a new, more agile approach to dynamic support for group learning where the strategy itself can adapt to the needs of the population of learners.

Conclusions and Current Work

In this paper we have laid an empirical foundation for a research agenda for a new generation of dynamic support interventions to improve collaborative learning, which we have termed *agile* support for collaborative learning. As we have demonstrated through an integration of results from four experimental studies, the effects of dynamic support vary based on the ability level of learners as well as the nature of the material itself. Human instructors are highly agile in their usage of complex interventions such as Academically Productive Talk along many dimensions, including selection of students, selection of facilitation moves, timing, and sequencing. Thus, we argue that achieving a higher level of agility is what is needed to move to the next stage—agility in terms of selection of students to target, selection of interaction strategies, and timing. Nevertheless, while the results presented in this article are compelling, it would be more compelling to examine the contrast between multiple different strategies within the same study. This will be important future work.

Agility comes with challenges from an experimental standpoint, however. As mentioned in the *architecture* discussion above, Bazaar's flexible approach to interactive script integration allows a variety of scripting paradigms to be implemented, with varying effects on the agents' internal and external validity. For example, specifying high priority and rigid constraints on macro-scripted actions, alongside low priority for dynamic feedback, produces an agent configuration with high internal experimental validity. In such a configuration, macro-script stages reliably occur at specified intervals, guaranteeing that each group of students interacting with instances of the agent engage with each stage of the script (and the associated learning opportunities) for the same amount of time. However, this comes with a loss of agility, and the potential for lost opportunities for natural collaborative conversation. The beginning of a new script phase may cut off an ongoing student conversation, or deny another component's chance to complete a follow-up move. On the other hand, if the dynamic components are configured to reserve more follow-up time after their behaviors are enacted (or the macro-script is configured to wait for a period of inactivity before preceding), there's greater opportunity for natural flow and resolution in student and agent interactions. This lends a greater external validity to the experience, but with greater variability in timing and experience between instances.

The technical approach presented in this article enables a wide variety of strategies to be implemented. The work presented in this paper provides the beginnings of the needed empirical foundation. However, we do not argue that the foundation provided here is sufficient. Rather, we offer this set of results as an argument in favour of a larger, more thorough and systematic investigation of the space of possibilities. We offer the publically available Bazaar architecture and the set of results presented here to the community, inviting further work from a broad and creative community of researchers working on intelligent support for group learning. While the statistical analysis technique used to estimate the effectiveness of a collaborative learning intervention is simple, we have demonstrated that it is highly accurate in separating effective from ineffective interventions. Because the approach is simple, it can be easily used by other researchers who take up the challenge to join the effort to fill out the space of results needed to work towards agile support for collaborative learning as a community of researchers.

Acknowledgments This work was supported in part by NSF grant SBE 0836012 to the Pittsburgh Science of Learning Center and NSF grant DRL-0835426.

References

Adamson, D., & Rosé, C. (2012). Coordinating multi-dimensional support in collaborative conversational agents. In *Proceedings of intelligent tutoring systems* (pp. 346–351). Berlin: Springer.

Adamson, D., Ashe, C., Jang, H., Yaron, D., & Rosé, C. (2013). Intensification of group knowledge exchange with academically productive talk agents. Proceedings of the 10th International Conference on Computer Supported Collaborative Learning, Madison Wisconsin, July 2013.

Adey, P., & Shayer, M. (1993). An exploration of long-term far-transfer effects following an extended intervention program in the high school science curriculum. *Cognition and Instruction, 11*(1), 1–29.

Ai, H., Kumar, R., Nguyen, D., Nagasunder, A., Rosé, C. P. (2010). Exploring the effectiveness of social capabilities and goal alignment in computer supported collaborative learning. In Proceedings of Intelligent Tutoring Systems, Lecture Notes in Computer Science volume 6095, pp 134–143.

Aronson, E., Blaney, N., Stephan, C., Sikes, J., & Snapp, M. (1978). *The jigsaw classroom*. Beverly Hills, CA: Sage Publications.

Azmitia, M., & Montgomery, R. (1993). Friendship, transactive dialogues, and the development of scientific reasoning. *Social Development, 2*, 202–221.

Berkowitz, M., & Gibbs, J. (1979). A Preliminary Manual for Coding Transactive Features of Dyadic Discussion. Unpublished manuscript, Marquette University.

Berkowitz, M., & Gibbs, J. (1983). Measuring the developmental features of moral discussion. *Merrill-Palmer Quarterly, 29*, 399–410.

Bill, V. L., Leer, M. N., Reams, L. E., & Resnick, L. B. (1992). From cupcakes to equations: the structure of discourse in a primary mathematics classroom. *Verbum, 1*(2), 63–85.

Chapin, S., & O’Connor, C. (2004). Project challenge: identifying and developing talent in mathematics within low-income urban schools Boston University School of Education Research Report No (vol. 1, pp. 1–6).

Chaudhuri, S., Kumar, R., Joshi, M., Terrell, E., Higgs, F., Aleven, V., et al. (2008). It’s not easy being green: supporting collaborative green design learning. Proc. Intelligent Tutoring Systems (ITS).

Chaudhuri, S., Kumar, R., Howley, I., & Rosé, C. P. (2009). Engaging collaborative learners with helping agents. In Proceedings of the 2009 conference on Artificial Intelligence in Education: Building Learning Systems that Care: From Knowledge Representation to Affective Modeling (pp. 365–372). IOS Press.

Clarke, S., Chen, G., Stainton, K., Katz, S., Greeno, J., Resnick, L., et al. (2013). The impact of CSCL beyond the online environment. *CSCL 2013 Conference Proceedings, 1*, 105–112.

de Lisi, R., & Golbeck, S. L. (1999). Implications of Piagetian Theory for peer learning. In A.M. O'Donnell & A. King (Eds.), *Cognitive perspectives on peer learning* (pp. 3–37). Mahwah, NJ: Lawerence Erlbaum Associates.

Dillenbourg, P. (2002). Over-scripting CSCL: The risks of blending collaborative learning with instructional design. In P. A. Kirschner (Ed.), *Three worlds of CSCL. Can we support CSCL* (pp. 61–91). Heerlen: Open Universiteit Nederland.

Dillenbourg, P., & Hong, F. (2008). The mechanics of CSCL macro scripts. *The International Journal of Computer-Supported Collaborative Learning, 3*(1), 5–23.

Dillenbourg, P., Baker, M., Blaye, A., & O’Malley, C. (1995). The evolution of research on collaborative learning. In E. Spada & P. Reiman (Eds.), *Learning in Humans and Machine: Towards an interdisciplinary learning science* (189–211).

Diziol, D., Walker, E., Rummel, N., & Koedinger, K. R. (2010). Using intelligent tutor technology to implement adaptive support for student collaboration. *Educational Psychology Review, 22*(1), 89–102.

Dyke, G., Adamson, D., Howley, I., & Rosé, C. P. (2013). Enhancing scientific reasoning and explanation skills with conversational agents. *IEEE Transactions on Learning Technologies, 6*(3), 240–247.

Erkens, G., & Janssen, J. (2008). Automatic coding of dialogue acts in collaboration protocols. *International Journal of Computer Supported Collaborative Learning, 3*, 447–470.

Fernando, S., & Stevenson, M. (2008). A semantic similarity approach to paraphrase detection. Computational Linguistics UK (CLUK 2008) 11th Annual Research Colloquium.

Graesser, A., VanLehn, K., the TRG, & the NLT. (2002). Why2 Report: Evaluation of Why/Atlas, Why/AutoTutor, and Accomplished Human Tutors on Learning Gains for Qualitative Physics Problems and Explanations. LRDC Tech Report, University of Pittsburgh.

Hmelo-Silver, C. E., & Barrows, H. S. (2006). Goals and strategies of a problem-based learning facilitator. *The Interdisciplinary Journal of Problem Based Learning, 1*(1), 21–39.

Hohenwarter, M., & Preiner, J. (2007). Dynamic mathematics with GeoGebra. *Journal of Online Mathematics and its Applications, 7*, article 1448.

Howley, I., Adamson, D., Dyke, G., Mayfiled, E., Beuth, J., & Rosé, C. P. (2012). Group composition and intelligent dialogue tutors for impacting students’ self-efficacy, ITS 2012 Proceedings of the 11th International conference on Intelligent Tutoring Systems, Lecture Notes in Computer Science volume 7315, Springer-Verlag, pp 551–556.

Howley, I., Kumar, R., Mayfield, E., Dyke, G., & Rosé, C. P. (2013). Gaining insights from sociolinguistic style analysis for redesign of conversational agent based support for collaborative learning. In D. Suthers, K. Lund, C. P. Rosé, C. Teplovs, N. Law (Eds.), *Productive multivocality in the analysis of group interactions*, edited volume, Springer.

Kirschner, F., Paas, F., & Kirschner, P. A. (2009). A cognitive load approach to collaborative learning: United brains for complex tasks. *Educational Psychology Review, 21*, 31–42.

Kobbe, L., Weinberger, A., Dillenbourg, P., Harrer, A., Hämäläinen, R., Häkkinen, P., et al. (2007). Specifying computer-supported collaboration scripts. *International Journal of Computer-Supported Collaborative Learning, 2*(2), 211–224.

Kollar, I., Fischer, F., & Hesse, F. W. (2006). Collaborative scripts—a conceptual analysis. *Educational Psychology Review, 18*(2), 159–185.

Kumar, R., & Rosé, C. P. (2011). Architecture for building conversational agents that support collaborative learning. *Learning Technologies, IEEE Transactions on, 4*(1), 21–34.

Kumar, R., Rosé, C. P., Wang, Y. C., Joshi, M., & Robinson, A. (2007). Tutorial dialogue as adaptive collaborative learning support. *Proceedings of the 2007 Conference on Artificial Intelligence in Education*, 383–390.

Kumar, R., Ai, H., Beuth, J., & Rosé, C. P. (2010). Socially-capable conversational tutors can be effective in collaborative learning situations. In Proceedings of Intelligent Tutoring Systems, Lecture Notes in Computer Science volume 6095, pp 156–164.

Lison, P. (2011). Multi-policy dialogue management. In Proceedings of the SIGDIAL 2011, pp. 294–300. Association for Computational Linguistics.

McLaren, B., Scheuer, O., De Laat, M., Hever, R., de Groot, R. & Rosé, C. P. (2007). Using machine learning techniques to analyze and support mediation of student E-discussions. Proceedings of Artificial Intelligence in Education, IOS Press, pp. 331–338.

Michaels, S., O'Connor, C., & Resnick, L. B. (2008). Deliberative discourse idealized and realized: accountable talk in the classroom and in civic life. *Studies in Philosophy and Education, 27*(4), 283–297.

Mihalcea, R., Corley, C., & Strapparava, C. (2006). Corpus-based and knowledge-based measures of text semantic similarity. Proceedings of the National Conference on Artificial Intelligence (AAAI 2006), Boston, Massachusetts, pp. 775–780.

Mu, J., Stegmann, K., Mayfield, E., Rosé, C. P., & Fischer, F. (2012). The ACODEA framework: developing segmentation and classification schemes for fully automatic analysis of online discussions. *International Journal of Computer Supported Collaborative Learning, 7*(2), 285–305.

Mühlpfordt, M., & Wessner, M. (2005). Explicit referencing in chat supports collaborative learning. In Proceedings of Computer Support for Collaborative Learning (CSCL '2005).

Rabe-Hesketh, S., & Skrondal, A. (2012). *Multilevel and longitudinal modeling using stata*. College Station, TX: Stata Press.

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). GLLAMM Manual. University of California, Berkely. U. C. Berkeley Division of Biostatistics Working Paper Series, Paper 160.

Resnick, L. B., Salmon, M., Zeitz, C. M., Wathen, S. H., & Holowchak, M. (1993). Reasoning in conversation. *Cognition and Instruction, 11*(3–4), 347–364.

Resnick, L. B., Asterhan, C. A., & Clarke, S. N. (2013). *Socializing intelligence through academic talk and dialogue*. Washington, DC: American Educational Reserach Association.

Rosé C., & VanLehn, K. (2005). An evaluation of a hybrid language understanding approach for robust selection of tutoring goals. *International Journal of Artificial Intelligence in Education, 15*(4), 325–355.

Rosé, C. P., Jordan, P., Ringenberg, M., Siler, S., VanLehn, K., & Weinstein, A. (2001). Interactive conceptual tutoring in Atlas-Andes. In *Proceedings of AI in Education 2001 Conference*, 151–153.

Rosé, C. P., Wang, Y. C., Cui, Y., Arguello, J., Stegmann, K., Weinberger, A., et al. (2008). Analyzing collaborative learning processes automatically: exploiting the advances of computational linguistics in computer-supported collaborative learning. *The International Journal of Computer-Supported Collaborative Learning, 3*(3), 237–271.

Scardamalia, M., & Bereiter, C. (1993). Technologies for knowledge-building discourse. *Communications of the ACM, 36*(5), 37–41.

Scardamalia, M., & Bereiter, C. (2006). Knowledge building: theory, pedagogy, and technology. The Cambridge handbook of the learning sciences, pp. 97–115.

Schwartz, D. (1998). The productive agency that drives collaborative learning. In P. Dillenbourg (Ed.), Collaborative learning: Cognitive and computational approaches. Bingley, UK: Emrald Group Publishing.

Soller, A., & Lesgold, A. (2000). *Modeling the process of collaborative learning. Proceedings of the International Workshop on New Technologies in Collaborative Learning*. Japan: Awaiji–Yumebutai.

Stahl, G., Koschmann, T., & Suthers, D. (2006). Computer-supported collaborative learning: An historical perspective. In R. K. Sawyer (Ed.), *Cambridge handbook of the learning* (pp. 409–426). Cambridge, UK: Cambridge University Press.

Suthers, D. (2006). Technology affordances for inter-subjective meaning making: a research agenda for CSCL. *International Journal of Computer Supported Collaborative Learning, 1*, 315–337.

Teasley, S. D. (1997). Talking about reasoning: How important is the peer in peer collaborations? In L. B. Resnick, C. Pontecorvo, & R. Saljo (Eds.), *Discourse, tools, and reasoning: Situated cognition and technologically supported environments*. Heidelberg: Springer.

Topping, K. J., & Trickey, S. (2007). Collaborative philosophical inquiry for schoolchildren: Cognitive gains at 2-year follow-up. *British Journal of Educational Psychology, 77*(4), 787–796.

Webb, N. M., & Palinscar, A. S. (1996). Group processes in the classroom. In D. C. Berliner & R. C. Calfee (Eds.), *Handbook of educational psychology* (pp. 841–873). New York: Prentice Hall.

Wecker, C., Fischer, F. (2007). Fading scripts in computer-supported collaborative learning: The role of distributed monitoring. CSCL'07 Proceedings of the 8th international conference on Computer supported Collaborative Learning, 764–772.

Wegerif, R., Mercer, N., & Dawes, L. (1999). From social interaction to individual reasoning: an empirical investigation of a possible socio-cultural model of cognitive development. *Learning and Instruction, 9*(6), 493–516.

Weinberger, A., & Fischer, F. (2006). A framework to analyze argumentative knowledge construction in computer-supported collaborative learning. *Computers & Education, 46*(1), 71–95.

Weinberger, A., Stegmann, K., Fischer, F., & Mandl, H. (2007). Scripting argumentative knowledge construction in computer-supported learning environments. In Scripting Computer-Supported Collaborative Learning, CSCL Book Series volume 6, chapter 6, pp 191–211.

Wiemer-Hastings, P., Graesser, A., Harter, D., & the Tutoring Research Group. (1998). The foundations and architecture of AutoTutor. In B. Goettl, H. Halff, C. Redfield, & V. Shute (Eds.), *Intelligent tutoring systems: 4th International Conference (ITS '98)* (pp. 334–343). Berlin: Springer.

Zinn, C., Moore, J. D., & Core, M. G. (2002). A 3-tier planning architecture for managing tutorial dialogue. In S. A. Cerri, G. Gouardères, & F. Paraguaçu (Eds.), *Proceedings of the Sixth International Conference on Intelligent Tutoring Systems, ITS 2002* (pp. 574–584). Berlin: Springer Verlag.