Using Automated Dialog Analysis to Assess Peer Tutoring and Trigger Effective Support

Erin Walker[1], Nikol Rummel[2], Kenneth R. Koedinger[1]

[1]Carnegie Mellon University, [2]Ruhr Universität Bochum
erin.a.walker@gmail.com, nikol.rummel@rub.de, koedinger@cmu.edu

Abstract. Intelligent tutors have the potential to be used in supporting learning from collaboration, but there are few results demonstrating their positive effects in this domain. One of the main challenges in automated support for collaboration is the machine classification of dialogue, giving the system an ability to know when and how to intervene. We have developed an automated detector of conceptual content that is used as a basis for providing adaptive prompts to peer tutors in high-school algebra. We conducted an after-school study with 61 participants where we compared this adaptive support to two nonadaptive support conditions, and found that adaptive prompts significantly increased conceptual help and peer tutor domain learning. The amount of conceptual help students gave, as determined by either human coding or machine classification, was predictive of learning. Thus, machine classification was effective both as a basis for feedback and predictor of success.

Keywords: intelligent tutoring, peer tutoring, adaptive collaboration support

1 Introduction

Computer-mediated collaborative learning activities have been demonstrated to improve student domain learning [1]. When students articulate their reasoning as part of interacting with others, they can engage in *beneficial cognitive processes*; they may reflect on misconceptions, elaborate on existing knowledge, and generate new knowledge [2]. However, without guidance, students may not collaborate in ways that lead them to benefit. One potential remediation is to add intelligent tutoring technologies that can assess the quality of collaboration as it occurs and provide targeted support. This support might lead students to engage in more beneficial cognitive processes as they try to collaborate better, causing an improvement in domain learning [3]. In a small number of studies, adaptive support for collaboration quality has indeed shown to be better than no support and nonadaptive support at increasing domain learning [e.g., 4]; in another small set, adaptive support has been shown to improve collaboration quality directly [e.g., 5]. However, there are no studies that have demonstrated an effect on both collaboration *and* learning. Thus, a causal link between adaptive support, improved collaboration, and learning has yet to be established. We explore that link by investigating three hypotheses (Figure 1): Adaptive support improves student learning (*H1a*), improves collaboration quality (*H1b*), and better collaboration quality relates to improved domain learning (*H1c*).

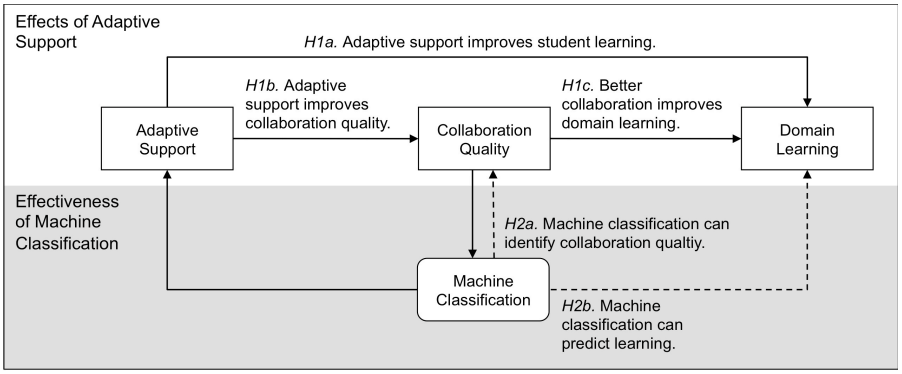

Figure 1. Hypotheses investigated. We explore the link between support, collaboration, and learning, and test how well our system classifies collaboration quality and learning.

One reason these hypotheses have not been fully explored might be that tutoring systems for collaborative learning are hard to construct. Collaboration quality is linked to properties of student dialogue, and to adaptively support this dialogue its properties need to be classified in real-time. In many existing systems, dialogue is assessed by having students self-classify their utterances [6]. For example, students may select a sentence starter like "I disagree, because…" in order to signal an instance of constructive conflict. However, students do not consistently select sentence starters that match the content of their statements, and therefore the inferences that the system makes can be inaccurate [7]. Consequently, researchers are starting to use machine classification to label student dialogue as it occurs, with goals ranging from determining the conversation topic to labeling a student's argument [8, 9]. As the quality of dialogue relates to whether students benefit from collaboration [10], improving our ability to automatically classify properties of student utterances would have two potential benefits: a) It would increase our ability to target support to those utterances, b) Given a relationship between collaboration and domain learning, it would enable us to predict learning based on the machine classification. Thus, this paper also investigates two technical hypotheses (Figure 1): Machine classification can identify collaboration quality (*H2a*) and predict domain learning (*H2b*).

We investigated these hypotheses in the context of an intelligent tutoring system for reciprocal peer tutoring in algebra, called the Adaptive Peer Tutoring Assistant (*APTA*). Reciprocal peer tutoring is a type of collaborative learning activity where two students of similar abilities take turns tutoring each other [11]. The goal of *APTA* is to improve peer tutors' domain learning by providing adaptive support for their help. In giving help, peer tutors benefit from reflecting and elaborating on their knowledge [2]. These beneficial cognitive processes can be triggered when peer tutors construct high quality help [10], but peer tutors tend to need support to do so. One type of high-quality help is conceptual help, in that it references domain concepts as part of a hint or explanation. For example, the phrase "You need to subtract the ax to get the two x's on the same side" would be considered conceptual. Fuchs and colleagues trained peer tutors to give conceptual help, and found that tutors that received this training learned more than tutors that did not [12]. In *APTA*, we follow up on these results by using a machine classification of conceptual help to support

peer tutors in giving more conceptual help. We discuss a study where we assessed whether *APTA* improved the conceptual content of peer tutor help and peer tutor domain learning. We then examine the effectiveness of *APTA* for classifying peer tutor conceptual help and serving as a basis for feedback. Although there are other aspects of student dialogue that are supported by our system and may relate to learning, given the length of this paper we focus here on conceptual help.

2 The Adaptive Peer Tutoring Assistant (APTA)

APTA is a peer tutoring addition to the Cognitive Tutor Algebra, a successful individual intelligent tutoring system for high school algebra [13]. In *APTA*, one student tutors another on literal equation solving problems where they are given an equation like "$ax + by = cx + dy$" and a prompt like, "Solve for x". Students are seated at different computers. Using menus, the tutee can select operations like "subtract from both sides" and then type in the term they would like to subtract. Peer tutors can see the tutee's actions, but are not able to perform actions in the problem themselves (C in Figure 2). Instead, they mark the tutee's actions right or wrong (D in Figure 2). Students discuss the problem in a chat window (A in Figure 2).

APTA provides peer tutors with prompts in the chat in order to encourage them to reflect and elaborate on their domain knowledge while providing more conceptual help. The computer prompts the peer tutor to reflect in the chat window (e.g., "Owl, think about the last help you gave. Why did you say that? Can you explain more?"), where "owl" is the peer tutor). These prompts are visible to both students (B in Figure 2), and might include positive reinforcement ("Good work! Hinting or explaining the reason for a step can help your partner learn how to do the step"), or tips for giving

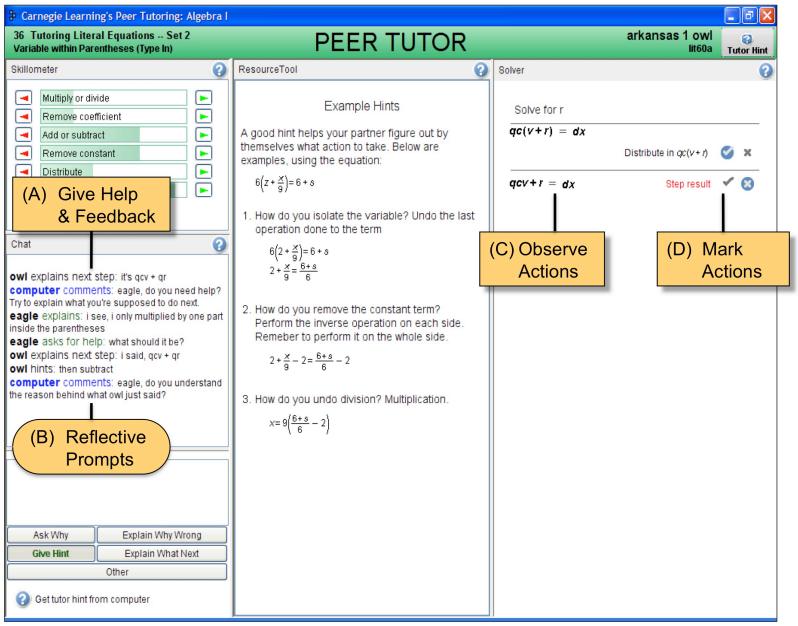

Figure 2. Peer tutor's interface in APTA. The peer tutor watches the tutee take problem-solving steps, and marks them correct or incorrect. The peer tutor helps the tutee in the chat window.

better help ("Owl, when helping, use examples or facts your partner already understands"). *APTA* incorporates prompts related to four different skills, namely (1) giving help when needed, (2) giving help targeting errors, (3) giving conceptual help, and (4) using the interface appropriately. Here, we focus on conceptual help.

Our assessment of whether students were giving conceptual elaborated help was based on an automated classification of student dialogue, described in [14]. We generated a baseline machine classifier for *conceptual content* using Taghelper Tools, state of the art text-classification technology designed for coding collaborative dialogue [9]. We then improved the accuracy of the classifier by adding three different types of domain features: problem-solving context (e.g., whether the tutee has just made an error), text substitutions (e.g., whether the peer tutor uses a domain-related word, like "add" or "isolate"), and substitution history (e.g., how many times in the past a given peer tutor has used a domain-related word). Training our automatic classification on previous study data, we achieved a kappa of 0.72 when compared to human raters. We expected accuracy to be lower when we deployed the system in the current study, given the change of population. Nevertheless, we used the machine classification of each dialogue utterance as part of a knowledge tracing model that assessed whether peer tutors knew how to give conceptual help, and, if not, triggered reflective prompts at relevant moments.

3 Method

As described in the introduction, we were interested in evaluating effects of adaptive support on the conceptual content of peer tutor help (*H1a*) and domain learning (*H1b*), with the hypothesis that conceptual help relates to learning (*H1c*). In a controlled study, we compared an adaptive support condition to two nonadaptive conditions. In the *real adaptive condition*, students received relevant prompts based on the automated assessment (using *APTA*). They were told that the prompts they received were adaptive ("The computer will watch you tutor and give you targeted advice when you need it based on how well you tutor"). In the *told adaptive condition,* we still told students that support was adaptive, using the above instructions. However, students were actually given nonadaptive support, where they received randomly selected prompts at moments when they would not have received the adaptive prompts. We ensured that the adaptive and random prompts appeared with the same frequency. In the *real nonadaptive condition,* students received the nonadaptive support and were told the support was not adaptive ("From time to time, the computer will give you a general tip chosen randomly from advice on good collaboration"). Including these two control conditions was an attempt at separating the cognitive effects of receiving support tailored to one's collaborative actions from the motivational effects of believing support is adaptive. If receiving adaptive support is indeed beneficial for improving help given by tutors, the *real adaptive condition* would have a better effect than the *told adaptive* and *real nonadaptive* conditions.

Participants were 130 high-school students (49 males, 81 females) from one high school, currently enrolled in Algebra 1, Geometry, or Algebra 2. While the literal equation solving unit was one that all students had (in theory) received instruction on

in Algebra 1, the teacher we were working with nevertheless identified it as a challenging unit for the students. The study was run at the high school, either immediately after school or on Saturdays. All students were paid 30 dollars for their participation, and as a result, appeared to be highly motivated during the study activities. Students participated in sessions of up to 9 students at a time. Each session was randomly assigned to one of the three conditions. Students came with partners that they had chosen, except for 4 students to whom the researchers then assigned partners. Within each pair students were randomly assigned to the role of tutee or tutor. Eight students worked alone and were not included in the analysis, leaving 122 students. For the purposes of this paper, we focus on peer tutor interaction and learning, and thus analyze data from 61 peer tutors.

Students first took a 20-minute domain pretest, and then spent 20 minutes working individually using the CTA to prepare for tutoring. They were then assigned either the tutor or tutee role. Students spent a total of 60 minutes in a tutoring phase, with one student tutoring another student. Finally, students took a 20-minute domain posttest. The pretests and posttests were counterbalanced, and contained conceptual and procedural items relating directly to the literal equation solving domain. To assess help quality and the accuracy of the automated classification, we human coded peer tutor help during tutoring for *conceptual content* by scoring whether each peer tutor utterance contained a reference to one or more domain concept. For example, "add ax to cancel out the $-ax$" and "cancel out the $–ax$" were conceptual, while "add ax" and "add ax so you can factor" were not. A total of 3105 utterances were made by peer tutors, and coded. To compute interrator reliability two independent raters coded 647 utterances separately, and achieved a kappa of 0.79.

4 Effects of Adaptive Support

To investigate the effects of condition on peer tutor learning (*H1a*), we conducted a one-way ANCOVA, with posttest score as the dependent measure, condition as a between subjects variable and pretest score as a covariate (see Table 1). Condition had a significant effect on posttest score ($F[2,57] = 4.47$, $p = 0.016$), and pretest was also significantly predictive of posttest score ($F[1, 57] = 33.24$, $p < 0.001$). Post-hoc contrasts revealed that students in the real adaptive condition learned significantly more than students in the real nonadaptive condition ($p = 0.019$) and marginally more than students in the told adaptive condition ($p = 0.077$), controlling for pretest. Overall, providing adaptive support led peer tutors to learn more, suggesting that the adaptive support triggered beneficial cognitive processes related to domain learning.

Next, we tested *H1b*, examining whether condition had an effect on conceptual content of tutor help. Here, we used negative binomial regression, because the outcome variable, conceptual content, was a count variable that was not normally distributed. We included two dummy coded condition variables in the regression, one representing the *told adaptive* condition and one representing the *real nonadaptive* condition, so that both could be compared to the *real adaptive* condition. We controlled for total help given by the peer tutor (all utterances that contained any domain information), which, using an ANOVA, was not significantly different

Table 1. Domain learning scores and amount of conceptual help.

	Pretest Score	Posttest Score	Conceptual Help	Total Help
Real Adaptive	0.27 (0.15)	0.39 (0.17)	4.16 (5.89)	26.00 (12.10)
Told Adaptive	0.24 (0.12)	0.27 (0.14)	1.77 (2.76)	32.55 (12.51)
Real Nonadaptive	0.29 (0.16)	0.28 (0.18)	3.15 (4.58)	29.85 (7.56)

between conditions ($F[1,58] = 1.82$, $p = 0.17$). The told adaptive condition was negatively related to the amount of conceptual help compared to the real adaptive condition ($\beta = -0.922$, $\chi 2(1, N = 61) = 3.976$, $p = 0.046$), and the real fixed condition was not significantly different from the real adaptive condition ($\beta = -0.310$, $\chi 2(1, N = 61) = 0.565$, $p = 0.452$). Essentially, when all else is held constant, the *real adaptive* condition is responsible for roughly 2.51 more instances of conceptual help per student than the *told adaptive* condition, and 1.36 more instances of conceptual help per student than the *real nonadaptive* condition. The total help was also related to the amount of conceptual help ($\beta = 0.039$, $\chi 2(1, N = 61) = 5.841$, $p = 0.016$).

Finally, we wanted to determine whether the conceptual help peer tutors gave was related to their domain learning (*H1c*). We conducted a linear regression with posttest score as the dependent measure, and conceptual content and pretest score as predictor variables. We also included the dummy coded condition variables to separate the overall effects of condition from the effects of conceptual content. We found that the conceptual content of help was marginally predictive of learning ($\beta = 0.199$, $t(60) = 1.95$, $p = 0.071$). As in our test of *H1a*, taking part in the actually adaptive condition significantly influenced learning compared to the *real nonadaptive* condition ($\beta = 0.308$, $t(60) = 2.64$, $p = 0.011$), and marginally influenced learning compared to the *told adaptive* condition ($\beta = 0.227$, $t(60) = 1.92$, $p = 0.060$). In sum, increased conceptual help partially mediated the effect of condition and learning, but there are likely other (yet unknown) interaction factors that had positive effects on learning.

5 Effectiveness of Machine Classification

We then examined how accurately our system assessed conceptual help. First, we compared the machine classification to the human codes on an utterance level (*H2a*). Table *2* displays the confusion matrix for the conceptual help codes. While the percent accuracy of the codes is 94%, with the vast majority of non-conceptual help correctly classified, *Cohen's kappa* is 0.53, as only 50% of the conceptual help instances were correctly classified. On the surface, this result would indicate that our classifier was less successful than we might have hoped. However, we can also explore the relationship between the human and computer coding on a student level, rather than on an utterance level, in order to assess more generally whether a given student has developed the ability to give conceptual help. The correlation between the human and machine count of instances of conceptual help *per student* was significant ($r[59] = 0.855$, $p < 0.001$), suggesting that the computer classification is overall accurate at determining whether students know how to give conceptual help. Thus, two goals of our classifier were met: a) It could identify instances of nonconceptual help in order to provide relevant support, and b) it could determine if a given student had the ability to give conceptual help by looking at the overall machine classifier count for that student. Further, *H2b* asked whether the machine classification of conceptual help could predict domain learning. Running the same regression as in Section 4, with posttest score as the dependent measure, and computer coded

Table 2. Confusion matrix for machine and human classification of conceptual content.

		Computer Codes	
		not conceptual	conceptual
Human Codes	not conceptual	2793	117
	conceptual	66	116

conceptual content, pretest score, and condition as predictor variables, we found that the computer classification was as predictive of student learning as the human classification ($\beta = 0.225$, $t(60) = 2.15$, $p = 0.036$). Using the machine classification of conceptual help, we can predict whether peer tutors will learn from the activity.

6 Discussion

In this paper, we described *APTA*, a system for adaptively supporting peer tutors in high school algebra. We discussed the component of *APTA* that detects peer tutor use of conceptual help and provides relevant prompts. We found that the prompts significantly increased peer tutor learning and the conceptual help peer tutors gave their partners. The amount of conceptual help given was marginally predictive of peer tutor learning, suggesting that there was indeed a causal link between the adaptive support provided, the increase in peer tutor conceptual help, and the increase in peer tutor learning. However, the relative weakness of the relationship between conceptual help and learning, and the differing pattern of results between the control conditions (the *real nonadaptive* condition learned the least but the *told adaptive* condition gave the least amount of conceptual help) suggested that there were other mediating factors at play. In fact, some of these factors may be relatively undetectable; the adaptive support, by prompting peer tutors to reflect at relevant moments, might increase their beneficial cognitive processes without having a tangible effect on the help they give. Further, it is likely that certain aspects of the peer tutor and tutee *interaction* (such as how much the tutee builds on peer tutor ideas) might have a positive effect on peer tutor learning. Nevertheless, this paper takes a step towards identifying the mechanisms by which adaptive support might lead to greater learning.

The second contribution of this paper is a technical one, examining the effectiveness of our machine classifier for conceptual help in this domain. On an utterance level the classifier was not as accurate as we might have hoped at positively identifying instances of conceptual help. However, on a practical level, the classifier was successful, and the support based on the classifier proved to be effective at improving conceptual help and domain learning. Indeed, accurate detection of non-conceptual help instances may be more valuable than accurate detection of conceptual help instances. Interestingly the classifier was accurate at assessing whether a given student was overall able to give conceptual help, and successfully predicted learning based on these classifications. This result suggests that these machine classifiers can function effectively as broader assessments of collaborative skill and domain learning.

This paper has focused on supporting conceptual help in a peer tutoring activity. However, we believe our results generalize to other collaborative learning activities, as conceptual elaboration and help exchange are key elements of collaboration in

general. One might also extend this technology to support a student in interacting with teachable agents or companion agents. By developing an understanding of how adaptive support can assess student collaboration, influence collaboration quality, and improve student domain learning, we can build powerful intelligent support systems for human-human and human-agent collaborative learning activities.

Acknowledgments. This work was supported by the Pittsburgh Science of Learning Center, NSF Grant #SBE-0836012. Thanks to Ruth Wylie, Ido Roll, Amy Ogan, Sean Walker, Carolyn Rosé, and the mathematics coordinator who made this possible.

References

1. Lou, Y., Abrami, P.C., d'Apollonia, S.: Small group and individual learning with technology: A meta-analysis. R. Ed. Res. 71(3), 449--521 (2001)
2. Ploetzner, R., Dillenbourg, P., Preier, M., Tram, D.: Learning by explaining to oneself and to others. In: Dillenbourg, P. (ed.) Collaborative Learning: Cognitive and Computational Approaches, pp. 103--121. Elsevier Science Publishers (1999)
3. Rummel, N., Weinberger, A.: New Challenges in CSCL: Towards Adaptive Script Support. In: Kanselaar, E., Jonker, V., Kirschner, P.A., Prins, F. (eds.) Proc. ICLS 2008, pp. 338--345. International Society of the Learning Sciences (2008)
4. Kumar, R., Rosé, C.P., Wang, Y.C., Joshi, M., Robinson, A.: Tutorial dialog as adaptive collaborative learning support. In: Luckin, R., Koedinger, K.R., Greer, J. (eds.) Proc. AIED 2007, pp. 383--390. IOS Press (2007)
5. Baghaei, N., Mitrovic, A., Irwin, W.: Supporting Collaborative Learning and Problem-Solving in a Constraint-Based CSCL Environment for UML Class Diagrams. IJCSCL. 2(2-3), pp. 159-190 (2007)
6. Soller, A., Martinez, A., Jermann, P., and Mu¨hlenbrock, M.: From mirroring to guiding: A review of state of the art technology for supporting collaborative learning. IJAIED, 15, 261–290. (2005)
7. Israel, J., Aiken, R.: Supporting collaborative learning with an intelligent web-based system, IJAIED, 17(1), 3-40. (2007)
8. Kumar, R., Rosé, C. P., Wang, Y. C., Joshi, M., Robinson, A.: Tutorial dialogue as adaptive collaborative learning support. AIED2007, 383-390. (2007)
9. Rosé, C., Wang, Y.-C., Cui, Y., Arguello, J., Stegmann, K., Weinberger, A., Fischer, F.: Analyzing collaborative learning processes automatically: Exploiting the advances of computational linguistics in computer-supported collaborative learning. IJCSCL. 3(3), 237-271. (2008)
10. Webb, N. M., & Mastergeorge, A. Promoting effective helping behavior in peer-directed groups. International Journal of Educational Research, 39, 73–97. (2003).
11. Dillenbourg, P., & Jermann, P. Designing integrative scripts. In: Fischer, F., Mandl, H., Haake, J. & Kollar, I. (Eds.), Scripting computer-supported communication of knowledge - cognitive, computational and educational perspectives, pp. 275–301 Springer (2007).
12. Fuchs, L., Fuchs, D., Hamlett, C., Phillips, N., Karns, K., & Dutka, S.: Enhancing students' helping behavior during peer-mediated instruction with conceptual mathematical explanations. The Elementary School Journal, 97(3), 223-249. (1997)
13. Koedinger, K., Anderson, J., Hadley, W., Mark, M.: Intelligent tutoring goes to school in the big city. IJAIED, 8, 30-43. (1997)
14. Walker, E., Walker, S., Rummel, N., & Koedinger, K.: Using Problem-Solving Context to Assess Help Quality in Computer-Mediated Peer Tutoring. ITS2010 (2010)