Trialog: How Peer Collaboration Helps Remediate Errors in an ITS

Robert G.M. Hausmann, Brett van de Sande, and Kurt VanLehn

Pittsburgh Science of Learning Center
Learning Research and Development Center, University of Pittsburgh
3939 O'Hara Street, Pittsburgh, Pa, 15260-5179
[bobhaus, bvds]@pitt.edu, vanlehn@cs.pitt.edu

Abstract

Many intelligent tutoring systems (ITSs) offer feedback and guidance through structured dialogs with their students, which often take the form of a sequence of hints. However, it is often difficult to replicate the complexity and responsiveness of human conversation with current natural language understanding and production technologies. Although ITSs reveal enough information to continue solving a problem, the conversations are not very engaging. To enhance engagement, the current study manipulated tutorial dialog by transforming them into a *trialog* by adding another student. Our intention was to advance the help offered by the system by putting students in a position to help each other, as well as make sense of the help offered by the ITS. The present paper attempts to show that conversations, either with the system or with a peer, are important design considerations when building an effective ITS.

Interactive Tutoring

Why should we expect that learning from interactive tutoring leads to stronger learning gains than non-interactive instruction? Intuitively, it makes sense that students should form a deeper understanding when they are asked a series of questions by the tutor in which they are expected to reply in natural language. But what does the empirical data suggest? The answer is mixed (VanLehn et al., 2007).

There is positive evidence to suggest that interactive tutoring leads to stronger learning gains than less interactive learning situations (such as reading). For instance, Evens and Michael developed an interactive tutoring environment in which medical students learn about the baroreceptor reflex, which is a part of the circulatory system that maintains a constant blood pressure across a variety of postures and conditions. Students using their system demonstrated larger learning gains than a control group that was asked to read a text written to match the content of the tutor (Rovick & Michael, 1992).

There is also convincing evidence to the contrary, that non-interactive learning situations are equally effective as interactive tutoring. For instance, Craig et al. (2006) contrasted an interactive tutor, called AutoTutor, with a modified version of the same system that only presented a didactic lecture. The lecture was non-interactive in the sense that students only watched the lecture. They were not required to respond to the system. The students demonstrated equal learning gains in both conditions. This result was surprising, especially because the two systems were closely matched for the content that was presented to the students.

Because there is convincing evidence, both for and against interactive tutoring, it appears that several variables are interacting, in complex ways, to produce the observed learning gains. One possible avenue for further exploration is to look at the dialogs and hints produced by the tutoring systems themselves.

Making sense of automatically generated hints

When an authority speaks, novices tend to listen. However, what if the authority does not make sense from the student's perspective? What are novices to do then? Consider the following hint provided by Andes, an intelligent tutoring system (ITS) for physics: "You should finish entering all of the useful given quantities in the problem. Why don't you work on entering the given value of the magnitude of the electric field at the region due to an unspecified agent." If you are unfamiliar with electric fields, you may not be aware that an electric field is a vector quantity, which is expressed as both a magnitude and a direction. Moreover, you may not be familiar with the way in which this particular tutoring system handles vectors and their magnitude representations. The bottom line is that this hint may not be helpful to the uninitiated.

Not surprisingly, novices sometimes find themselves in this unfortunate position while using an ITS. For whatever reason, the help emanating from an ITS might not be all that helpful. When this happens, the student soon learns either: a.) to drill down to the terminal (or "bottom-out") hint, or b.) randomly input slightly different entries until the system marks it as correct. Because the help system is unable to sufficiently guide students around their current impasse, these "help abusers" can miss potentially important learning opportunities.

What can be done to rectify this situation? One approach is to observe students' interactions with the help system and modify them to promote positive help-seeking behaviors. For instance, Baker, et al. (2006) designed an ITS to detect when students abuse the help system. When students begin to game the system, they are given negative feedback in a creative way. An animated agent, which is a puppy, becomes angry when it detects that the student is heavily abusing the help. When the system first detects that students are gaming the system, they are shunted off to a remedial exercise that targets the material that the student initially attempted to bypass.

On the other end of the spectrum are students who fail to realize that they should ask for help in the first place. In an effort to remediate this situation, Roll et al. (2007) modified a version of the Cognitive Tutor to help students assess their competency and understanding of the target domain. They created the "Help-seeking Support Environment" by modifying the wording of the tutorial hints, as well as adding prompts that hinted at the meta-cognitive level. They also presented a short classroom lesson that stressed productive help-seeking behaviors. The results suggested that the modified tutoring environment had an impact on the students' declarative understanding of the help-seeking process.

Another approach to the problem of making sense of automatically generated help by a tutoring system is to modify the interactions themselves. Take, for instance, a learning situation in which college nursing students (i.e., unskilled tutors with domain expertise) were asked to tutor a group of eighth graders on the topic of the circulatory system (Chi, Siler, Jeong, Yamauchi, & Hausmann, 2001). The tutors were very effective in producing large learning gains. In a follow-up experiment, the tutors were asked not to give away any information. Instead, they were told to elicit the information from the students. This manipulation of the tutorial interaction proved to be a success because the students in the second experiment learned just as much as the students in the first.

In a similar vein, the present experiment attempted to increase the help efficacy by modifying the tutorial interactions. Toward that end, we interjected another student into the conversation. The reason for adding a peer is to harness the beneficial effects of collaboration, which are covered in the next section.

Collaboration and the Andes physics tutor

Like interactive tutoring systems, collaboration has been found to be effective under certain circumstances, yet it fails to produce robust learning differences in other situations (Hill, 1982). Collaboration is hypothesized to be effective when the interaction is structured in some way. Efforts to scaffold collaboration include using sentence openers (Soller, 2004), including collaboration scripts (Rummel & Spada, 2005), and tutoring collaboration itself (Walker, Koedinger, McLaren, & Rummel, 2006). Systems that guide or script collaboration vary in the explicitness of their intervention. For example, requiring a sentence opener for each turn is strongly structured and explicit, whereas modeling the behavior and periodically reminding the dyad to engage in a certain type of interaction is much less structured and explicit (Hausmann, 2006).

The Andes physics tutor does not script collaboration *per se*. Instead, student conversations were structured in two ways. The first is somewhat implicit, while the second was more explicit. The implicit structuring of collaborative interactions was realized via the step-based nature of the tutoring. Andes is a step-based tutoring system, in the sense that it offers feedback and help on each step of a problem. Users can ask Andes for help on an incorrect entry, as well as what to do next. The help is offered in a series of hints that decrease in abstraction, with top-level hints that are very general and abstract, to bottom-out hints that tell the student explicitly which action to take (see Fig. 1). The implication is that the conversations that students have during problem solving were implicitly structured to focus on individual problem-solving steps.

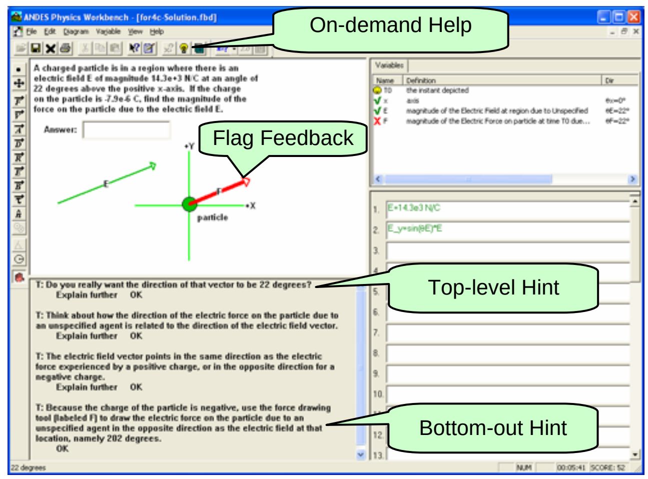

Figure 1. A screen shot of Andes.

To explicitly structure the dialogs, we chose a form of conversation that has been shown to be effective in previous studies on collaborative learning. Explanatory activities were chosen because they have been shown to be useful for both individual (Chi, Bassok, Lewis, Reimann, & Glaser, 1989) and collaborative learning (Coleman, 1998; Ploetzner, Dillenbourg, Praier, & Traum, 1999). To structure the conversations in the present experiment, we provided instructions that described how collaborative explanations should unfold, as well as provided an example of a hypothetical dyad producing a "joint explanation." We also designed a set of prompts to remind students to provide explanations while studying worked-out examples.

The reason for structuring the dialog was to help the students solve the physics problems in an effective way. Although great care was taken in designing the help system and hints for each of the problems in Andes, there are unfortunate times when the hints fail to help the student take the next correct step in solving the problem. Should such an event arise, the student has little recourse. He or she can repeatedly request bottom-out hints until the entire solution is revealed. At this point, the student has unfortunately lost ownership of the solution and may learn little from the experience.

Alternatively the student might ask a peer for guidance. Because peer help typically occurs in classroom life, we decided to investigate collaborative problem solving in an ITS more systematically. Therefore, we conducted the following laboratory experiment, in which we contrasted students who were asked to use the system alone with those who were asked to work collaboratively.

Method

Participants

Thirty-nine undergraduates (N = 39), enrolled in a second semester physics course, were randomly assigned to one of two experimental conditions: self-explanation (individuals; n = 11) or joint-explanation (dyads; n = 14). Volunteers were recruited from several sections of a second-semester physics course, which covered Electricity and Magnetism. Participants were recruited during the third week of the semester, with the intention that the experimental materials would coincide with their introduction in the actual physics course. The participants were paid $10 per hour. To ensure that the participants' motivation remained high during the entire two-hour session, they were offered an incentive of an additional $10 for doing well on the tests. All of the students received the bonus.

Materials

The materials developed for this experiment were adapted from an earlier experiment (Hausmann & VanLehn, 2007). The domain selected for this experiment was electro-dynamics with a focus on the definition of the electric field, which is expressed by the vector equation: $\mathbf{F} = q\mathbf{E}$. This particular topic is typically covered within the first few weeks of a second-semester physics course. Thus, it is an important concept for students to learn because it represents their first exposure to the idea that a field can exert a force on a body.

To instruct the participants, several materials were developed. Four electrodynamics problems were created. These problems are representative of typical problems found at the end of a chapter in a traditional physics textbook. The problems covered a variety of topics, including the definition of the electric field; Newton's first and second law, the weight law, and several kinematics equations. Each of the four problems was implemented in Andes. The first problem served as a warm-up problem because none of the students had any prior experience with the Andes user interface.

In addition to the problems, three examples were created in collaboration with two physics instructors at the U.S. Naval Academy. The examples contained a voice-over narration of an expert solving the problems, and they were structured such that they were isomorphic to the immediately preceding problem.

Procedure

The procedure consisted of several activities. The first activity was to watch a short, introductory video on the Andes user interface. Afterwards, the participants read instructions on how to produce explanations, including an example. Next, participants were asked to use Andes to solve a warm-up problem. The experimenter was available to answer any user-interface questions. He was not, however, allowed to give away any domain-specific information. During the problem solving, the student had access to the flag feedback (correct/incorrect), the hint sequences, and an equation cheat sheet. Once the student submitted a final answer, she then watched and explained an example of an expert solution of an isomorphic problem. The example solutions were broken down into steps, and at the conclusion of each step the student was prompted to explain (either individually or jointly). Once the explanation was complete, the participant clicked a button to go onto the next step. Only the cover story and given values differed between the problem-solving and example problems. The students alternated between solving problems and studying examples until all four problems were solved and all three examples were studied, or until two hours elapsed.

Instead of using a traditional *pretest-intervention-posttest* design, the current study employed a micro-genetic approach where the density of observations is increased to reveal a fine-grained profile of changing behavior (Siegler & Crowley, 1991).

Measures

Several dependent measures were used to assess problem-solving performance. We used the log files generated by Andes to count the number of errors the students made while solving problems, the number of hints requested, and specifically the number of bottom-out hints requested.

Results

The results are presented in three sections. The first section presents the differential use of the hints. The second section presents an analysis of the error rates, which are then divided into shallow and deep errors. Finally, the third section briefly analyzes a segment of dialog taken from a dyad's solution.

On-demand help: Hints and bottom-out hints

Before delving into the error and hint analyses, it should be noted that more dyads correctly solved the final problem than the individuals, $\chi^2(1, N = 25) = 4.81$, $p < .03$. This means that, because the experiment was capped at two hours, some individuals (8/11 = 72.7%) were unable to finish all of the problems (see Fig. 2).

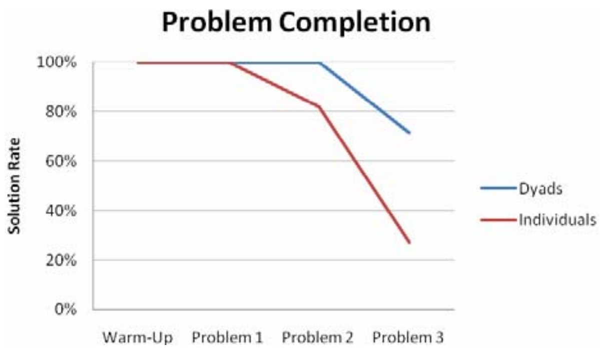

Figure 2. Percentage of students completing each problem.

Symptomatic of their inability to complete the entire problem set was the individuals' reliance on the help system to get them through the problems. The individuals (M = 94.09, SD = 51.92) asked for nearly twice as many hints as the dyads (M = 48.57, SD = 39.59). This relationship was statistically reliable with a large effect size, $F(1, 23) = 6.20$, $p = .02$, $d = 1.05$.

Moreover, the same pattern of results held for the number of bottom-out hint requests. The individuals (M = 13.91, SD = 14.36) requested nearly three times as many bottom-out hints as the dyads (M = 4.71, SD = 7.32). Again, this difference was statistically reliable with a large effect size, $F(1, 23) = 4.34$, $p < .05$, $d = .88$.

Errors: Deep versus shallow

Why did the individuals rely so heavily on the hints? We tested two potential hypotheses. First, perhaps dyads entered fewer deep errors; therefore, they did not require as deep of hinting to remediate their errors. Second, the dyads may have turned to each other for guidance, instead of asking the tutoring system for assistance. Notice that these explanations are not mutually exclusive. It could be the case that the dyads made fewer deep, conceptual errors *and* they remediate their errors in dialog.

To test the first hypothesis, we identified and hand-coded N = 928 errors committed by the individuals and dyads. We coded the errors according to their depth. Slips, typos, and user-interface errors were coded as *shallow*, while conceptual errors were labeled *deep*. Collapsing across the experimental conditions, there were n = 443 shallow errors and n = 475 deep errors. The remaining ten errors did not fit into either category and were placed into a *miscellaneous* category.

Overall, there were slight differences between the two conditions in terms of the total number of errors made, while controlling for the total number of student entries. Across all four problems, the error rate for the dyads (M = .33, SD = .08) was lower than the error rate observed for the individuals (M = .40, SD = .10). The difference in error rates was marginally significant with a medium to large effect size, $F(1, 23) = 3.23$, $p = .09$, $d = .75$.

In terms of the depth of errors, the dyads (M = .17, SD = .04) and individuals (M = .17, SD = .05) demonstrated an equivalent number of shallow errors. There were, however, reliable differences in terms of the deep error rates. Dyads (M = .15, SD = .08) committed fewer deep errors per entry than the individuals (M = .20, SD = .09). Using a repeated-measures ANOVA, there was a statically reliable condition-by-problem interaction, $F(3, 63) = 2.72$, $p = .05$, $\eta_p^2 = .114$. When we analyzed the number of deep errors as a function of problem, it became evident that the dyads started out with proportionally fewer deep errors, $F(1, 21) = 6.40$, $p = .02$, $d = 1.13$. This difference gradually diminished over the course of the experiment (see Fig. 3).

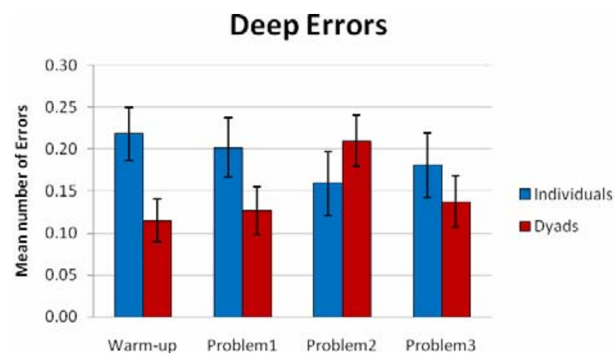

Figure 3. Mean number of deep errors per problem.

This set of results suggests that the first hypothesis was partially supported. The dyads' use of hints may have reflected the fact that they did not need them as frequently as the individuals because they committed fewer deep errors, at least early in the experiment.

Trialog: Collaborative error remediation

Because we have not yet coded the entire corpus of verbal protocols, we sampled the dialogs to detect any patterns of interaction (Stahl, 2004). While a complete test of the second hypothesis is ongoing, this section presents suggestive evidence of a representative exchange between a pair of students while they were engaged in problem solving in the context of a hint.

In the following episode (see Table 1), two students are attempting to write the following equation: Fe = abs(qe)*E. Before equations can be written in Andes, all variables must be defined. Because Andes told the dyad that "E" was not yet defined, they set about attempting to draw the vector representation of "E." Unfortunately, Andes turned the vector red (line: 1). Because the E-field vector contained an error, Andes did not yet recognize it as a legitimate variable, so when the dyad entered their equation, Andes told them that "E" was not recognized (line: 2). This was somewhat unintelligible to the pair (line: 3 & 4), so they asked Andes for another hint (line: 5). This hint was helpful because they both recognized the error (lines: 6 & 7) and fixed it (line: 10).

Next, they went back and hit "enter" on their equation. Andes continued to complain because it said, "Units are inconsistent" (line: 12). Ziggy determined the correct units and announced it to his partner (line: 13). It appears that Ozzy did not immediately understand Ziggy because evidence of his understanding came a bit later (line: 16). Once they both were on the same page, they simultaneously stated the equation (lines: 17 & 18) and solved for the sought quantity (line: 20).

Table 1. A trialog between Andes and Dyad 24.

Line	Speaker	Statement
1	Ozzy	Oh, it didn't like it.
2	Andes	*The variable "E" is not used in any solution I know of.*
3	Ziggy	Aren't we looking for "E" though?
4	Ozzy	What's wrong with that? [Request next hint.]
5	Andes	*Is the electron the source of the electric field you are defining?*
6	Ozzy	The electron
7	Ziggy	Oh, the electron's not the source.
8	Ozzy	Okay
9	Ziggy	So we just need to put it as "unspecified." [Andes turns the vector green.]
10	Ozzy	That would make sense.
11		[Types: E = abs(qe)*Fe and Andes turns the equation red.]
12	Ozzy	[Reads hint] "Units are inconsistent"
13	Ziggy	F equals Newtons per Colomb.
14	Ozzy	Huh.
15	Ziggy	Newtons per Colomb
16	Ozzy	Oh wait, Newtons. Yeah, it should be
17	Ziggy	/F divided by /
18	Ozzy	/F divided by /
19		[Types: E = Fe / abs(qe) and Andes turns the equation green.]
20	Ozzy	Solve for "E"

There are a few interesting features of this particular trialog. First, Andes is not always immediately understood by the pair. Instead, the dyad asked Andes for clarification, either in the form of another hint in the sequence, or another hint altogether. Alternatively, one partner may understand what Andes is saying and announce his insight to his partner, as was the case in the vague "Units are inconsistent" hint. The second interesting feature was that, while the dyad was not shy about asking Andes for hints, they did not abuse the help. Instead, they thought carefully about what it was trying to say. They did not immediately drill down to the bottom-out hint. Instead, they chose to remediate their errors with high-level hints and lots of interaction between themselves.

Discussion

During the early development of the field, one of the open research questions was, "Should students use ITSs alone or in small groups?" In some cases, the limited number of computers in a classroom governed whether students worked alone or in pairs. The learning that resulted from non-systematic pairing of students was equivalent to the students who did not use the tutor (Anderson, Corbett, Koedinger, & Pelletier, 1995).

When we remove the pragmatic limitations of conducting evaluations in the real world, the outlook for collaborative use of an ITS is more promising. For example, the Sherlock system was originally designed to train Air Force technicians. It was then modified to capitalize on the strengths of the reflection dialogs that occurred after the students used the training system (Lesgold, Katz, Greenberg, Hughes, & Eggan, 1992). Using the system in a peer group was advantageous for the technicians because they were in a position to help their partner, as well as to direct their partner's use of Sherlock (Katz & O'Donnell, 1999).

Therefore, instead of asking, "Do dyads benefit more from using an ITS than individuals?" a more appropriate question might be, "What are the types of student conversations that facilitate learning in an ITS?" Lesgold and Katz would suggest the answer is "follow-up reflection dialogs." Our suggested answer is that the trialog should focus on remediating errors on steps.

To support this claim, we looked at several different measures of learning, most of which demonstrated a fairly coherent and consistent pattern of results. The dyads requested fewer hints and bottom-out hints than the individuals. To explain why, we analyzed the types of errors that were made while solving problems. At the largest grain size, the dyads produced fewer errors per entry than the individuals. At a more fine-grained analysis, the dyads demonstrated fewer deep errors than the individuals. This suggests that the dyads did not require as many hints because they were not making as many deep errors.

When help was requested, dyads were less likely to request a bottom-out hint. Again, this could be due to the lower number of deep errors, or it might also be explained by the dyads' use of the hint sequences. For instance, an excerpt from one of the dyads demonstrated that their conversation was focused on making sense of the Andes hint. This segment of speech illustrated the collaborative nature of the trialogs. Andes presented a suggestion to the dyad, but it becomes the dyad's mission to make sense of the hints. Dyads may have been able to avoid asking for a bottom-out hint because they reasoned through the higher-level hints.

Although ITSs do not yet employ full natural-language understanding and generation, this may be a potential opportunity for enhancing learning. If the conversations between students can be scripted in such a way that they encourage students to work together to understand the help emanating from the tutoring system, then there may be an added advantage of engaging students in the active construction of their own knowledge and understanding. The implication for the design of an ITS is a consideration for the inclusion of support mechanisms for error-remediation dialogs.

Acknowledgements

This work was supported by the Pittsburgh Science of Learning Center, which is funded by the National Science Foundation award number SBE-0354420. The authors are deeply indebted to Robert Shelby and Donald Treacy for their assistance in developing the instructional materials, and to Jim Greeno's *Interactive Analysis* group for their constructive criticism of an earlier version of this paper.

References

Anderson, J. R., Corbett, A. T., Koedinger, K., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *The Journal of the Learning Sciences, 4*, 167-207.

Baker, R. S. J. d., Corbett, A. T., Koedinger, K. R., Evenson, E., Roll, I., Wagner, A. Z., et al. (2006). Adapting to when students game an intelligent tutoring system. In M. Ikeda, K. D. Ashley & T.-W. Chan (Eds.), *Proceedings of the 8th International Conference on Intelligent Tutoring Systems* (pp. 392-401). Berlin: Springer-Verlag.

Chi, M. T. H., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science, 13*, 145-182.

Chi, M. T. H., Siler, S. A., Jeong, H., Yamauchi, T., & Hausmann, R. G. (2001). Learning from human tutoring. *Cognitive Science, 25*(4), 471-533.

Coleman, E. B. (1998). Using explanatory knowledge during collaborative problem solving in science. *The Journal of the Learning Sciences*, *7*(3&4), 387-427.

Craig, S. D., Sullins, J., Witherspoon, A., & Gholson, B. (2006). The deep-level reasoning questions effect: The role of dialog and deep-level reasoning questions during vicarious learning. *Cognition and Instruction, 24*(4), 565-591.

Hausmann, R. G. M. (2006). Why do elaborative dialogs lead to effective problem solving and deep learning? In R. Sun & N. Miyake (Eds.), *28th Annual Meeting of the Cognitive Science Society* (pp. 1465-1469). Vancouver, B.C.: Sheridan Printing.

Hausmann, R. G. M., & VanLehn, K. (2007). Explaining self-explaining: A contrast between content and generation. In R. Luckin, K. R. Koedinger & J. Greer (Eds.), *Artificial intelligence in education: Building technology rich learning contexts that work* (Vol. 158, pp. 417-424). Amsterdam: IOS Press.

Hill, G. W. (1982). Group versus individual performance: Are N + 1 heads better than one? *Psychological Bulletin, 91*(3), 517-539.

Katz, S., & O'Donnell, G. (1999). The cognitive skill of coaching collaboration. In C. M. Hoadley & J. Roschelle (Eds.), *Proceedings of the 1999 conference on Computer support for collaborative learning* (pp. 291-299). Stanford, CA.

Lesgold, A., Katz, S., Greenberg, L., Hughes, E., & Eggan, G. (1992). Extensions of intelligent tutoring paradigms instructional models in computer-based learning environments. In S. Dijkstra, H. Krammer & J. van Merrienboer (Eds.), *Instructional Models in Computer-Based Learning Environments.* (pp. 291-311). Berlin: Springer-Verlag.

Ploetzner, R., Dillenbourg, P., Praier, M., & Traum, D. (1999). Learning by explaining to oneself and to others. In P. Dillenbourg (Ed.), *Collaborative learning: Cognitive and computational approaches* (pp. 103-121). Oxford: Elsevier.

Roll, I., Aleven, V., McLaren, B. M., & Koedinger, K. R. (2007). Can help seeking be tutored? Searching for the secret sauce of metacognitive tutoring. In R. Luckin, K. R. Koedinger & J. Greer (Eds.), *Artificial intelligence in education: Building technology rich learning contexts that work* (Vol. 158, pp. 203-210). Amsterdam: IOS Press.

Rovick, A. A., & Michael, J. A. (1992). The predictions table: A tool for assessing students' knowledge. *American Journal of Physiology*, *263*(6), S33-S36.

Rummel, N., & Spada, H. (2005). Learning to collaborate: An instructional approach to promoting collaborative problem solving in computer-mediated settings. *Journal of the Learning Sciences*, *14*(2), 201-241.

Siegler, R. S., & Crowley, K. (1991). The microgenetic method: A direct means for studying cognitive development. *American Psychologist, 46*, 606-620.

Soller, A. (2004). Computational modeling and analysis of knowledge sharing in collaborative distance learning. *User Modeling and User-Adapted Interaction, 14*, 351–381.

Stahl, G. (2004). Building collaborative knowing: Elements of a social theory of learning. In J.-W. Strijbos, P. A. Kirschner & R. Martens (Eds.), *What we know about CSCL in higher education* (pp. 53-85). Amsterdam, NL: Kluwer.

VanLehn, K., Graesser, A. C., Jackson, G. T., Jordan, P., Olney, A., & Rose, C. P. (2007). When are tutorial dialogues more effective than reading? *Cognitive Science, 31*(1), 3-62.

Walker, E., Koedinger, K., McLaren, B., & Rummel, N. (2006). Cognitive tutors as research platforms: Extending an established tutoring system for collaborative and metacognitive experimentation. In M. Ikeda, K. D. Ashley & T.-W. Chan (Eds.), *Intelligent Tutoring Systems* (pp. 207-216). Berlin: Springer.