

# Evaluating the Effect of Feedback from a CSCL Problem Solving Environment on Learning, Interaction, and Perceived Interdependence

Gahgene Gweon, Carolyn P. Rosé, Emil Albright, Yue Cui,  
Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh PA 15213,  
{gkg,cp3a,ealbrigh,ycui}@andrew.cmu.edu

**Abstract:** In this paper, we explore the effect of the form of feedback offered by a computer supported collaborative learning (CSCL) environment on the roles that students see themselves as taking and that their behavior reflects. We do this by experimentally contrasting collaboration in two feedback configurations, one which is identical to the state-of-the-art in intelligent tutoring technology (Immediate Feedback), and one which is based on a long line of investigation of the use of worked out examples for instruction (Delayed Feedback). While our conclusions remain tentative due to the small sample size, the data reveal a consistent gender by condition interaction pattern across questionnaire, test, and discourse data in which male students prefer and benefit more from collaboration in the Immediate Feedback condition where they are more likely to take on the role of a help provider rather than a help receiver while the patterns is the opposite for females.

## Introduction

In this paper we present an empirical investigation of issues related to the design of a collaborative problem solving environment that builds on prior work related to the development of intelligent tutoring technology for individual learning. We argue that the state-of-the-art in intelligent tutoring technology has been optimized for success in an individual learning scenario, and that many interaction design issues may need to be revisited in order to achieve success in a collaborative learning setting. In this paper we specifically investigate issues related to the timing of feedback from the intelligent tutoring environment. Immediate feedback involving indications of correct versus incorrect problem solving actions and hints on demand or unsolicited hints during problem solving are the hallmark of state-of-the-art intelligent tutoring technology. However, it is not clear whether such feedback from the intelligent tutoring environment will be helpful or harmful in a collaborative learning setting. This paper investigates the hypothesis that the typical state-of-the-art form of intelligent tutoring feedback interferes with collaborative learning because it can be treated as a replacement for the interaction between students that collaborative learning is meant to encourage. Because math has been a very popular domain for exploration in the intelligent tutoring community, we conducted our explorations in that domain. In particular, we selected fraction arithmetic as a unit of material because of its importance and difficulty for middle school students, which is our target student population.

For decades a wide range of social and cognitive benefits have been extensively documented in connection with collaborative learning. Based on Piaget's foundational work (Piaget 1985), one can argue that a major cognitive benefit of collaborative learning is that when students bring differing perspectives to a problem solving situation, the interaction causes the participants to consider questions that might not have occurred to them otherwise. This stimulus could cause them to identify gaps in their understanding, which they would then be in a position to address. This type of cognitive conflict has the potential to lead to productive shifts in student understanding. Related to this notion, other cognitive benefits of collaborative learning focus on the benefits of engaging in teaching behaviors, especially deep explanation (Webb, Nemer, & Zunita 2002). Other work in the computer supported collaborative learning community demonstrates that interventions that enhance argumentative knowledge construction, in which students are encouraged to make their differences in opinion explicit in collaborative discussion, enhances the acquisition of multi-perspective knowledge (Fischer, et. al 2002). Furthermore, based on Vygotsky's seminal work (Vygotsky 1978), we know that when students who have different strengths and weaknesses work together, they can provide support for each other that allows them to solve problems that would be just beyond their reach if they were working alone. This makes it possible for them to participate in a wider range of hands-on learning experiences. It is in connection with this Vygotskian model of collaborative learning that we see a conflict with the design of feedback, sometimes called scaffolding, that is the hallmark of the state-of-the-art in intelligent tutoring technology and is based on the same principles, and thus designed to meet the same needs. Our hypothesis predicts that the presence of typical intelligent tutoring style feedback in a collaborative problem solving

environment will reduce the amount of interaction students will engage in. Furthermore, a reduction in collaborative interaction may then lead to a reduction in the exchange of alternative perspectives on problem solving, thus also interfering with the benefits of collaboration from the Piagetian perspective.

While these cognitive benefits of collaborative learning are valuable, they are not the only positive effect of collaborative learning. In fact the social benefits of collaborative learning may be even more valuable for fostering a productive classroom environment. By encouraging a sense of positive interdependence between students, where students see themselves both as offering help and as receiving needed help from others, collaborative learning has been used as a form of social engineering for addressing conflict in multi-ethnic, inner-city classrooms (Sharan 1980). Some examples of documented social benefits of successful collaborative learning interactions include increases in acceptance and liking of others from different backgrounds, identification with and commitment to participation in a learning community, improvements in motivation, and aptitude towards long term learning (Sharan 1980). These social benefits of collaborative learning are closely connected with the Vygotskian foundations of collaborative learning because the positive interdependence that is fostered is related to the exchange of support, or scaffolding, that we hypothesize will be replaced with the scaffolding offered by the environment where typical intelligent tutoring technology is used.

In our experimental approach, we seek to balance concerns related to internal and external validity by running our experiment as a controlled experiment in a realistic setting (i.e., within a pair of real classrooms using material from their actual curriculum). Classroom settings present experimental challenges because there are always more factors beyond our control than in a lab setting. The two classes we worked with were small, having only 30 students in total across the two sets of students. Thus, with small such sample size, we struggle with issues related to statistical power. To increase our certainty in the conclusions we draw from our data, we consider only significant ( $p < .05$ ) and marginally significant ( $p < .1$ ) effects, making a distinction between these two in terms of certainty. Furthermore, we rely on a form of triangulation, to verify that we see a consistent story across multiple channels of data. We investigate the impact of this experimental manipulation on perceptions about the collaboration revealed by a questionnaire, evidence of learning from tests and quizzes, and a qualitative analysis of the collaborative problem solving process from coded chat logs collected during the collaborative problem solving sessions. We measure process oriented outcome measures such as observed help offered and observed help received through analysis of chat behavior. Furthermore, we measure perceived help offered and perceived help received by means of a questionnaire. We also measure cognitive benefits of collaboration, such as learning as measured by pre to post-test gains in domain knowledge. The data do not support the strong form of our initial hypothesis. Rather, we find a consistent gender by condition interaction across all forms of data we collect in which male students prefer and benefit more from the Immediate Feedback condition while female students prefer and benefit more from the Delayed Feedback condition. While the results we present are not conclusive enough to warrant offering concrete design principles yet, they raise important questions to resolve in subsequent work.

## **Infrastructure for Supporting Collaborative Problem Solving**

In this section we discuss the experimental infrastructure used to conduct our investigation, both in terms of the technology we used and in how we set up the lab where the students worked. Because of its tremendous effectiveness for individual learning with technology, we are planning to build our eventual collaborative learning infrastructure on the foundation of Cognitive Tutors (Koedinger, et. al 1997). Other development work related to supporting collaborative learning in connection with Cognitive Tutors is found in (Walker 2005). Our current infrastructure was built with the Cognitive Tutor Authoring Tools, which support quick authoring of Cognitive Tutor style problem solving systems. As mentioned, the purpose of our study is to explore issues relating the design of the problem solving feedback offered by the environment during collaborative problem solving. The infrastructure used in this study is a simple extension of the typical structured problem solving interfaces that are characteristic of Cognitive Tutors and other tutors in the model tracing tutor tradition. This infrastructure is designed to support experimentation with alternative feedback designs keeping all other aspects of the student's experience constant across conditions.

In our study we are contrasting two designs for feedback from the environment, which we refer to as Immediate feedback and Delayed feedback. These alternative feedback paradigms have been experimentally contrasted in individual learning settings in the past (e.g., Bjork 1994, Nathan 1998). Typically, immediate feedback consists of what is called flag feedback, which signals to students after each problem solving action

whether it was correct or not, and hints on demand, which are typically arranged in hint sequences, beginning with less directive hints and ending with more directive hints. In a delayed feedback setting, flag feedback is typically withheld so that students must use their own self-monitoring skills to detect their errors. Furthermore, hints may be withheld altogether or changed in nature so as not to be as focused narrowly on the correct solution path so that students have a greater responsibility for keeping themselves on track. In our study, both flag feedback and hints were withheld from students in the Delayed feedback condition. Instead, when students decided that their solution was complete, they submitted the solution and then were presented with a fully worked out version of the problem, with some explanation about how the solution was constructed. In order to control for information access between conditions, the instructional content in the explanation was constructed by concatenating the content encoded in the hints that students had access to in the Immediate feedback condition.

Based on prior work, we know there are trade-offs between immediate and delayed feedback for individual learning, especially regarding efficiency and retention (Bjork 1994, Nathan 1998). Studies have shown that immediate feedback is more efficient because students are never allowed to stray too far from the correct solution path. Therefore, a shorter amount of time is required to solve each problem, and in practice, students solve more problems (Corbett & Anderson 2002). Yet, other studies show that students get a deeper understanding of material in a delayed feedback setting since they have time to reflect on their errors and also that they have the opportunity to develop self-monitoring skills. This was shown in cognitive tasks such as learning genetics (Lee 1992) as well as in motor tasks such as learning arm movement motions (Schmidt & Bjork 1992). Most state of the art intelligent tutoring systems such as Cognitive Tutors have adopted an immediate feedback approach because in practice, the greater efficiency leads to higher learning gains in an individual learning scenario because of the relatively large numbers of problems students are able to work through. However, we conjecture that the optimal problem solving feedback design in a collaborative learning setting may be different.

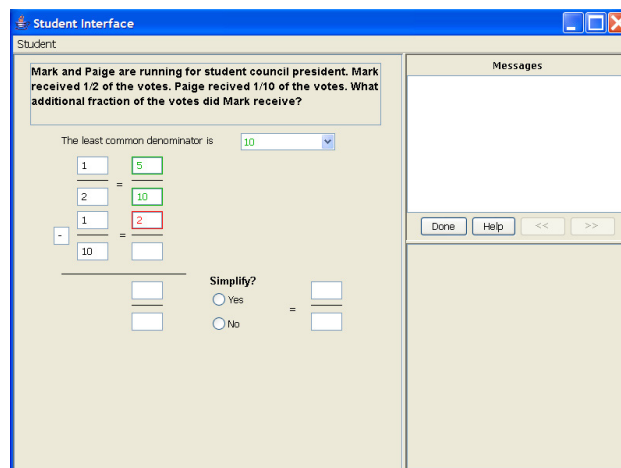


Figure 1. Problem solving interface for Immediate Condition.

As the students worked in the lab session, their computer's display was composed of two panels that were next to one another. In the panel on the left hand side of the screen, displayed in Figure 1, was the problem solving interface. Using RealVNC's Virtual Network Computing (<http://www.realvnc.com/>), this panel was shared between the screens of the respective computers of a collaborating pair so that they were both free to contribute to the evolving joint solution. In the panel on the right hand side of the screen was an MSN messenger window in which students could chat about their joint problem solving. The arrangement of the lab in which our study was conducted was such that each student was sitting at his own computer in such a way that collaborating pairs could not easily talk face-to-face since in all cases there was a row of desks with computers in between that student's row and the row where the partner student was sitting. The students did not know who their partner was or where they were seated. The purpose of this arrangement was to encourage communication through the chat interface so that it could easily be recorded and eventually processed on line during collaboration.

## Method

### Experimental Design and Procedure

We designed an experiment to test the hypothesis that if students are working together in an environment in which they can obtain immediate feedback and help from the environment that is always correct, they would be less likely to turn to each other for help and feedback. This hypothesis predicts that in an environment with this form of feedback students would give and receive less help, would perceive less help given and received, and would benefit less from the collaboration. In the control condition, students get immediate feedback from the cognitive tutor (Immediate Feedback condition), whereas the experimental condition students get delayed feedback (Delayed Feedback condition). Both of these feedback configurations are described in detail in the previous section.

The experimental procedure extended over 4 school days, with the experimental manipulation taking place during days two (i.e., Lab Day 1) and three (i.e., Lab Day 2), which we refer to as the first and second lab day since the students worked together in pairs in a computer lab at their school. The fourth day of the experiment was separated from the third day of the experiment by a weekend. Because our study is a within subject manipulation, we used two different units of material, each of which was experienced by each pair in only one condition or another so that we could distinguish learning resulting from work in one condition from learning resulting from work in the other condition. The two units were fraction addition and subtraction (AddSub) and fraction multiplication and division (MultDiv). We counter-balanced the order of the units and conditions in order to control for ordering effects as displayed in Table 1.

Table 1. The experimental setup

	Pairs	Lab Day 1	Lab Day 2
Class 1	1~4	AddSub, Imm	MultDiv, Delay
	5~8	MultDiv, Delay	AddSub, Imm
Class 2	9~11	AddSub, Delay	MultDiv, Imm
	12~15	MultDiv, Imm	AddSub, Delay

On the first day of the four day study, students took a pretest, which lasted for about 30 minutes, to assess how much they knew about the subject matter. We also provided a short collaboration training manual, where the teacher gave an example of good collaboration conversation. In addition, pairs of students were teamed by the instructor. Teams remained stable throughout the experiment. The students were instructed that the teams would compete for a small prize at the end of the study based on how much they learned and how many problems they were able to solve together correctly. The second and third days were lab days in which the students worked with their partner on one of the units in one of the conditions. On each lab day they worked through a different unit in a different condition from what they were in the previous day. Each lab session lasted for 35 minutes. At the end of each lab period, the students took a short quiz, which lasted about 10 minutes. At the end of the first lab day only, students additionally filled out a short questionnaire to assess their perceived help received, perceived help offered, and perceived benefit of the collaboration. On the fourth experiment day, which was two days after the last lab day, they took a post test, which was designed to be isomorphic to the pre test and was used for the purpose of assessing retention of the material.

### Subjects and Materials

Thirty sixth grade students from a suburban elementary school participated in the study. The students were from 2 different classes taught by the same teacher, with 16 students in the first class and 14 students in the second class. Students were arranged into arbitrary pairs by their instructor. Students were not told who their partner was. We had a mixture of mixed-ability and homogenous ability pairs. Furthermore, out of 15 pairs who participated in the study, 12 of them were mixed gender pairs, 2 of them were all female pairs, and one of them was an all male pair. Because only a small number of pairs were homogeneous gender pairs, we cannot draw any conclusions from this data about the relative merits of mixed gender versus homogeneous gender pairs. Furthermore, we cannot distinguish between gender effects that are specific to mixed gender pairs, versus gender effects that are independent of group composition.

The materials for the experiment consisted of the following:

- A mathematics tutoring program. The two mathematics chapters were fraction addition & subtraction and fraction division & multiplication.
- 2 extensive isomorphic tests (Test A and Test B) were designed for use as the pre-test and the post-test. These tests each consisted of 16 near transfer and 8 far transfer problems, balanced between the two units of material. Likewise, we had Quiz A and Quiz B, which were designed to be isomorphic to a subset of the pre/post tests. Thus, quizzes are shorter versions of the tests, administered after each lab day. Thus, we were able to use pre to post test gains as a measure of retention (since there was a two day lag between the last lab day and the post-test day).
- Questionnaire. As a subjective assessment of socially oriented variables, we used a questionnaire with 8 questions related to perceived problem solving competence of self and partner, perceived benefit, perceived help received, and perceived help provided. Each question consisted of a statement such as “The other student depended on me for information or help to solve problems.” and an 11 point scale ranging from -5, labeled “strongly disagree”, to +5, labeled “strongly agree”.

## Results

### Questionnaire

We began our analysis by investigating the socially oriented variables measured by means of the questionnaire, specifically perceived problem solving competence of self and partner, perceived benefit, perceived help received, and perceived help provided. Neither of our experimental conditions maximized all of these outcome variables for both genders. Instead we see a consistent gender by condition interaction across perceived benefit, perceived help received, and perceived help offered, although it is only significant in some cases and marginal in others. Specifically, in the Delay condition boys rated themselves as offering more help and receiving less as well as benefiting less, whereas the pattern was the opposite for girls, although the effect was not as strong.

Consistent with prior work investigating the well known gender gap in math achievement for middle school children, we found a main effect of gender whereby boys rated themselves on the questionnaire as being more competent problem solvers  $F(1,29) = 5.01, p < .05$ , effect size .7 s.d., although there was no significant difference in grade so far in the class reported by their teacher  $F(1,29) = 0.46, p = \text{n.s.}$  There was, however, a significant difference in pretest score whereby boys scored higher than girls  $F(1, 29) = 6.13, p < .05$ , effect size 1.2 s.d., thus demonstrating that they came into the experiment with more prior knowledge about the specific material covered. In terms of perceived benefit from the collaboration, boys rated themselves as benefiting significantly less than girls did  $F(1,29) = 2.15, p < .05$ . As mentioned, there was a significant interaction with condition such that the difference is only significant in the Delay condition  $F(1,29) = 4.63, p < .05$ , effect size 2.5 s.d. This effect did not seem to be related to the relatively higher pretest scores of boys since there was no significant correlation between perceived benefit and either the pretest score of the student or that of their partner. Related to perceived help provided we also found a significant gender by condition interaction  $F(1,29) = 4.84, p < .05$ . Specifically, girls’ ratings of the extent to which they offered help was significantly lower than that of boys, but only in the Delay condition. There was a corresponding marginal gender by condition interaction  $F(1,29) = 2.62, p = .1$  whereby girls’ ratings of the extent to which they received help were higher in the Delay condition, whereas the opposite was the case for boys.

### Learning Gains

The learning gains analysis is consistent with the interaction between gender and condition observed on the questionnaire and offers some weak evidence in favor of the Delay condition on learning overall. There was no measurable gain on far transfer items either within conditions or over the whole population, thus we suspect that the far transfer items may have been too difficult for these students, and we consider only learning on near transfer items for the remainder of our analyses to distinguish between conditions. We focus first on immediate learning. For our measure of immediate learning, we measured learning gains that occurred locally within each single lab session. Recall that the pre and post test were more extensive than the two quizzes, but contained a section that was isomorphic to the quizzes in order to enable a consistent measure of growth in understanding of the material over the 4 days of the experiment. The post test for each lab session was the quiz administered on the day of the session. For the first lab session, the pretest was the score on the subset of the pretest from day 1 of the study that was isomorphic to the quizzes. The pretest for the second lab day was the quiz score from the first lab day. We only considered data from the 12 out of 15 pairs for which both students were present for the pretest and both lab days.

For this analysis we used an ANCOVA model with post-test score as the dependent variable, condition, pair nested within condition, unit of material, time point, and gender as independent variables, and pre-test score as the covariate. The purpose of this ANCOVA design was to control for all of the factors that may have accounted for performance differences on the test, such as which units of material the students had been exposed to, when the test was administered, and gender (since we observed gender effects in the data). There was a marginal effect of pair on learning gains  $F(11,32) = 1.94$ ,  $p = .07$ , but no effect of unit of material (i.e., AddSub versus MultDiv) or time point (i.e., lab session 1 versus lab session 2). We see a marginal crossover interaction between gender and condition on near transfer items such that there was a trend for girls to learn more on average than boys in the Delay condition, and for boys to learn more on average than girls in the Immediate condition  $F(1,32) = 3.43$ ,  $p = .07$ . While it was true that boys came in to the experiment with higher pretest scores, we do not find a significant or even marginal aptitude-treatment interaction that might provide an alternative explanation for the gender by condition interaction on learning.

Because the strongest evidence presented thus far is for the Delay condition to be bad for boys, and only marginally significant evidence in favor of the benefit of the Delay condition for girls, one might argue that the data suggest that the most reasonable implication of these results would be to choose the Immediate feedback condition for all students. However, on the retention test, there was only a significant pre to post test gain in the Delay condition. For this analysis, because each student learned each unit of material in a different condition, in order to measure learning per condition, it is necessary to separate the test questions into subsets related to each unit. If a student learned the AddSub unit in the Delayed Feedback condition, then that student's pre and post test score for the Delayed Feedback condition would be the score on the part of the pre and post tests that were related to AddSub, and the corresponding portions of the tests related to MultDiv would be that student's pre and post test scores for the Immediate Feedback condition. We dropped from the analysis data from segments of material that students were absent for. One student did not take the post test, and 3 students were absent on the second lab day, one in the Immediate Feedback condition and 2 in the Delayed Feedback condition. Thus we have 56 pairs of scores, 29 for the Immediate condition and 27 for the Delay condition.

We computed the significance of the pre to post test difference using 2-tailed paired t-tests. Note that this analysis controls for pair effects and gender effects since all comparisons are for scores pertaining to an individual student. As mentioned, the difference was significant in the case of the Delay condition  $t(26) = 1.58$ ,  $p < .05$ , but not in the case of the Immediate condition  $t(28) = 2.27$ ,  $p = .12$ . This is consistent with the findings from other studies in that delayed feedback fosters deeper understanding of the material and thus would be beneficial for retention of the material.

### **Process Analysis**

The student chat logs contain rich data on how the collaborative problem solving process transpired. We conducted a qualitative analysis of the conversational data recorded from MSN messenger in order to illuminate the findings from the tests and questionnaire data discussed above. Based on the analysis of the questionnaire data, we expected to find that boys offered more help in the Delayed Feedback condition but received more help in the Immediate Feedback condition, and that the opposite would be the case for girls. However, we found on the one hand some surprising relationships between chat behavior and questionnaire data and on the other hand more straightforward relationships between patterns in the chat data and how much students learned. Specifically, we find that the condition where students offer more help is the condition where they perceive more benefit and learn more.

In order to make the sometimes cryptic statements of students clearer during our analysis, and also to provide an objective reference point for segmenting the dialogue into meaningful units, we merged the logfile data recorded by the tutoring software with the chat logs recorded with MSN messenger using time stamps for alignment. We then segmented the data into episodes using the log files from the tutoring software as an objective guide. Each episode was meant to include conversation pertaining to a single problem solving step as reified by the structured problem solving interface. All entries in the log files recorded by the tutoring software refer to the step the action is associated with as well as any hints or other feedback provided by the tutoring software.

We approached the design of our coding scheme with some focal questions in mind. For example, we wanted to investigate how many times each student requested help in each condition. Furthermore, we wondered how their partners responded to their help requests. A preliminary cursory analysis of the MSN messenger logs revealed that frequently students requested help but did not receive any verbal response from their partner. We also

observed signs of frustration between students and some cases where students explicitly refused to help one another. Because our focal questions all pertain to issues related to help seeking and help provision, we designed a coarse grained coding scheme to identify the regions of the integrated logfiles where this help seeking and help providing behavior is found. In the future we may code additional types of behaviors or make finer grained distinctions. Our current coding scheme has 5 mutually exclusive categories, namely (R) Requests received, (P) Help Provision, (N) No Response, (C) Can't Help, and (D) Deny Help. Along with the "other" category, which indicates that a contribution does not contain either help seeking or help providing behavior, these codes can be taken to be exhaustive. A sample of coded dialogue is found in Table 1 where the second and third columns contain the assigned codes. Each column is associated with a single conversational participant.

The first type of conversational action we coded were Help Requests (R). Help Requests are conversational contributions such as asking for help on problem solving, asking an explicit question about the domain content, and expressing confusion or frustration. Not all questions were coded as Requests. For example, there were frequent episodes where students discussed coordination issues such as whether the other student wanted to go next, or if it was their turn, and these questions were not coded as help requests for the purpose of addressing our research questions. Adjacent to each coded help request, in the column associated with the partner student, we coded four types of responses. Help provisions (P) are actions that attempt to provide support or substantive information related to the other student's request, regardless of the quality of this information. These actions are attempts to move toward resolving the problem. Can't help statements (C) are responses where the other student indicates that he or she cannot provide help because he or she doesn't know what to do either. Deny help (D) statements are where the other student responds in such a way that it is clear that he or she knows the answer but refuses to stop to help the other student. For example, "Ask [the teacher], I understand it" or "Hold on [and the other student proceeds to solve the problem and never comes back to answer the original question]" are type D statements. And finally, no response (N) are statements where the other student ignores help requests completely.

Table 2: Example Coded Conversation. Note that for simplicity, portions of the integrated logfile related to the interaction with the problem solving interface have been removed.

Line	S23	S24	speaker: content
92			s24: ur turn
93			s23: k
94	R	P95	s23: is it 1/20?
95			s24: no it is 4/20
96	R	P97	s23: y?
97			s24: cause to get 5 to 20 you need to multiply it by 4 and what you do to the bottom you must do to the top
98			s23: oooooo
99			s23: IM SO SRY
100			s23: :\$
101			s24: thats ok
102	R	P103	s23: i feel like a dope
103			s24: :D
105			s23: your turn
106			s24: k
107	P108	R	s24: you have to subtract right
108			s23: yea
110			s24: k
113	P114	R	s24: do you want to do the simplify it
114			s23: Sure
137	C138	R	s24: whats wrong with it
138			s23: idk [I don't know]

Each log file was coded separately by 2 coders who then met and resolved all conflicts. Using this consensus coding, we then tabulated the number of occurrences of each code in each condition associated with each gender. An example of one such interaction is displayed in Table 2. Here students take turns working out parts of a math problem (line 92, 105, 103). When help is requested, the other student provides an answer with some explanation (line 97). Such successful interactions in which students benefit from help received from their partner and also see themselves as contributing to the success of their partner promote feelings of positive interdependence between students (Sharan 1980). In Table 3 we display the average counts of actions within a single problem solving session. We tabulated the codes from the perspective of each student so that for each student we obtained a count for help requests made during the associated session as well as help requests received. We also noted how many problems were solved by that student working with his or her partner during the associated lab session as well as how many conversational segments there were in the integrated logfile.

Table 3 Average numbers (and standard deviation) of coded categories per session. Note that statistical comparisons in the body of the paper are presented both in terms of raw numbers and proportions.

	Males Immediate (7)	Females Immediate (9)	Males Delay (8)	Females Delay (6)
Problems Solved	10.0 (7.19)	6.0 (5.6)	5.0 (2.4)	4.7 (2.8)
Segments	21.7 (11.6)	17.1 (11.2)	15.1 (4.4)	16.8 (3.6)
(R) Requests Received	5.6 (3.3)	2.4 (1.2)	4.6 (3.5)	6.5 (4.5)
(P) Help Provision	3.3 (1.9)	0.6 (0.7)	2.0 (1.9)	3.3 (3.1)
(N) No Response	1.7 (1.4)	1.3 (1.1)	2.2 (1.6)	2.2 (3.9)
(C) Can't Help	0.3 (0.5)	0.6 (0.7)	0.1 (0.4)	0.7 (0.8)
(D) Deny Help	0.3 (0.8)	0 (0)	0.3 (0.7)	0.3 (0.8)
R Given	2.6 (0.8)	4.8 (3.4)	5.4 (4.4)	5.5 (3.5)
P Received	1.0 (1.0)	2.3 (2.3)	2.5 (3.7)	2.7 (1.8)
N Received	1.1 (.9)	1.8 (1.4)	2.1 (3.4)	2.3 (1.5)
C Received	0.4 (.8)	0.4 (0.5)	0.5 (0.7)	0.2 (0.4)
D Received	0 (0)	0.2 (0.7)	0.25 (0.7)	0.3 (0.8)

As a manipulation check, after we tabulated the number of occurrences of each code in each integrated log file, we first checked to see whether there was a significant effect of condition on patterns of occurrence of the codes. For this analysis, each count pertained to a single lab session, but we used data from both lab sessions. There was a marginal main effect of condition on number of problems solved  $F(1,44) = 3.49, p = .07$ , and a significant main effect of condition on number of segments  $F(1, 44) = 9.45, p < .005$ , with no interaction with gender. The larger average number of problems solved and larger average number of segments was found in the Immediate Feedback condition. This is to be expected based on prior findings that immediate feedback increases problem solving efficiency. While there was a significantly larger number of conversational segments in the integrated logs from the Immediate Feedback condition, the proportion of segments that contained a help request was not stable across conditions. Thus, there was no significant main effect of condition on raw numbers of either help requests received or offered. There was, however, a significant gender by condition interaction on raw number of requests received  $F(1,42) = 4.79, p < .05$ , and a marginal gender by condition interaction on both help requests given and help requests received when the raw counts are normalized by number of segments:  $F(1,42) = 3.62, p = .06$  and  $F(1,42) = 3.10, p = .09$  respectively. In all cases there was no significant or marginal gender effect except in the Immediate feedback condition, where males received more requests than females as well as participating in a higher proportion of discourse segments in which they received a request than females did. In contrast, females participated in a higher proportion of segments in which they made requests than males did.

Taking into consideration that the majority of collaborating pairs were mixed gender pairs, this analysis suggests that in the Immediate feedback condition, we find an asymmetric collaboration pattern in which males appear as the help providers and females appear as the help receivers. To further investigate this finding, we compared counts of response types across conditions, normalized by number of requests. Data from transcripts where no requests were received were dropped from this analysis. There was a significant main effect of condition on number of Can't Help responses such that a larger proportion of requests were met with a Can't Help response in the Immediate Feedback condition than in the Delayed Feedback condition, with no interaction with gender  $F(1,42) = 4.86, p < .05$ , effect size 1.5 standard deviations. This suggests that the nature of help requests may have been different in the two conditions. Our coarse grained coding of the collaborative behavior does not allow us to further address the question of what caused this difference at this time.

For the other three response types, we see a significant gender by condition interaction but no main effect of condition: Help Provision  $F(1,40) = 4.84, p < .05$ ; Deny Help  $F(1,40) = 3.96, p < .05$ ; No Response  $F(1,40) = 4.91, p < .05$ . For girls, the proportion of Help Provision and Deny Help responses is lower in the Immediate Feedback condition than in the Delayed Feedback condition, but higher for No Response responses. The pattern is almost the opposite for boys, where proportion of Deny Help responses remains stable between conditions, but the proportion of No Response responses is lower in the Immediate Feedback condition than the Delayed Feedback condition, and the proportion of Help Provision responses is higher in the Immediate Feedback condition than the Delayed Feedback condition. Thus, the asymmetric collaboration pattern reverses directions between conditions when we examine responses to help requests. Whereas girls offer more help in the Delayed Feedback condition, boys offer more help in the Immediate Feedback condition.

We examined relationships between patterns of occurrence of those codes in the collaborative process and the quantitative social and cognitive outcome measures coming from the questionnaire data and the tests and quizzes. These findings are described in the following two sections. Our purpose has been to inform the design for a collaborative learning environment that will enhance positive interdependence between students as well as facilitating learning. However, based on the questionnaire data, neither of our conditions consistently maximized all three of our socially oriented dependent variables, namely perceived benefit, perceived help received, and perceived help offered. The surprising finding is that it appears that girls perceive themselves as benefiting more and receiving more help in the condition in which they are actually offering more help, and conversely, boys see themselves as receiving more help and benefit in the condition in which they are offering more help. Specifically, what we found was a male preference for the Immediate feedback condition and a female preference for the Delayed feedback condition such that girls perceived themselves as receiving more help and more benefit in the Delayed Feedback condition, whereas the pattern was the opposite for boys. In terms of perceived help offered, there was no difference between how girls and boys rated themselves in the Immediate Feedback condition, but girls rated themselves as offering significantly less help in the Delayed Feedback condition than boys did. As mentioned, what we observed based on our corpus analysis is that girls responded to a higher proportion of help requests with a substantive answer in the Delayed Feedback condition, whereas boys responded to a higher proportion of help requests with a substantive answer in the Immediate Feedback condition.

One possible explanation for perceiving more help where one is in fact offering more help is that the act of offering help is an instructionally beneficial activity, and then when students engage in this activity, they perceive themselves as receiving help and benefit because they are learning. Recall that in the learning gains analysis reported above with the quantitative analysis, we observed that girls learned more in the Delayed Feedback condition where we see them offering more help, whereas boys learned more in the Immediate Feedback condition where we see them offering more help. As further evidence of this connection we see a significant correlation between total number of Help Provision responses and learning when we compute a multiple regression with pretest score and number of Help Provision responses as independent variables and posttest score as the dependent variable ( $R^2 = .84, p = .001, N = 30$ ) and a significant gender by condition interaction on total number of Help Provision occurrences that mirrors the earlier analysis with respect to proportion of Help Provision responses  $F(1,26) = 7.79, p = .01$ . A Bonferroni posthoc analysis reveals a marginal difference between number of Help provision statements made by girls in the Delayed Feedback condition and that in the Immediate Feedback condition (effect size .89 standard deviations) and a marginal difference between number of Help provision statements made by boys in the Immediate Feedback condition and by girls in the Immediate Feedback condition (effect size .86 s.d.).

## Conclusion

We have investigated the hypothesis that the presence of typical intelligent tutoring system style feedback in a collaborative problem solving interface would interfere with collaboration and dampen its positive effects. While our data do not support the strong version of this hypothesis, we are left with the challenge of reconciling the dichotomous needs and preferences of girls and boys. Further experimentation is required to identify a satisfactory solution. One obvious follow-up study that we plan to run is to replicate the design from this study except using only homogeneous gender pairs rather than mixed gender pairs. This would allow us to separate gender preferences that are specific to mixed-gender pairs from those that are more generally gender based. Further analysis of the data from this investigation might yield additional insights that would allow us to identify other possible ways of reconciling the different needs and preferences of girls and boys. For example, while we have evidence that our experimental manipulation lead to increases in productive behavior for learning in one condition for boys and the other for girls, we do not know why they responded more positively to different conditions. There may be deeper differences in the interaction styles characteristic of each feedback condition that are obscured by our coarse grained analysis of the data. We believe a deeper analysis of our conversational data would yield new insights.

This work was supported by National Science Foundation grant number IERI REC-043779.

## References

- Anderson, J.R., Boyle, C.F., Corbett, A., Lewis, M.W. (1990). Cognitive modeling and intelligent tutoring. *Artificial Intelligence*, 42, 7-49.
- Bjork, R. A. (1994). Memory and metameory considerations in the training of human beings. In J. Metcalfe and A. Shimamura (Eds). *Metacognition: Knowing about knowing*. Cambridge, MA: MIT Press. 1994, 185-205
- Corbett, A. T., and Anderson, J. R. (2002). Locus of feedback control in computer-base tutoring: impact on learning rate, achievement and attitudes. In *Proceedings of CHI 2002*, ACM 2002, 245-252
- Elbers, E., De Hann, M. Dialogic Learning in the Multi-Ethnic Classroom. *Dialogic Learning: Shifting Perspectives to learning, instruction and teaching*.
- Fischer, F., Bruhn, J., Gruesel, C & Mandl, H. (2002). Fostering collaborative knowledge construction with visualization tools. *Learning and Instruction*, 12, 213–232.
- Koedinger, K. J., Anderson, R. J., Hadley, W.H., Mark, M.A. (1997). Intelligent Tutoring Goes to School in the Big City. *International Journal of Artificial Intelligence in Education*. 8, 30-43
- Lee, A. Y. (1992). Using tutoring systems to study learning: An application of HyperCard. *Behavior Research Methods, Instruments, & Computers*, 24 (2), 205-212
- Nathan, M. J. (1998). Knowledge and situational feedback in a learning environment for algebra story problem solving. *Interactive Learning Environments*, 161-180
- Palinscar, A.S., Brown, A. L. (1984). Reciprocal teaching of comprehension-fostering and comprehension-monitoring activities. *Cognition and Instruction 1* p.117-175.
- Piaget, J. (1985). *The equilibrium of cognitive structures: the central problem of intellectual development*, Chicago University Press.
- Rummel, N., Spada, H., Caspar, F., Ophoff, J. G., Schornstein, K. (2003). Instructional support for computer-mediated collaboration – results from process analyses. In *Proc. CSCL 2003*, 199-208.
- Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: common principles in three paradigms suggest new concepts for training. *Psychological Science*, 3 (4), 207-217
- Sharan, S. (1980). Cooperative Learning in Small Groups: Recent methods and Effects on Achievement, Attitudes, and Ethnic Relations. *Review of Educational Research*, Vol 50, No. 2, 241-271
- Vygotsky, L.S. (1978). *Mind and society: The development of higher mental processes*. Cambridge, MA: Harvard University Press
- Walker, E. (2005). Mutual Peer Tutoring: A Collaborative addition to the algebra-1 Cognitive Tutors. *Young Researchers Track at AIED 2005*
- Webb, N., Nemer, K., Zuniga, S. (2002). Short Circuits or Superconductors? Effects of Group Composition on High-Achieving Students' Science Assessment Performance, *American Educational Research Journal*, 39, 4, 943-989.