Comparing Triggering Policies for Social Behaviors

Rohit Kumar, Carolyn P. Rosé
Language Technologies Institute, Carnegie Mellon University
Gates Hillman Center, 5000 Forbes Avenue,
Pittsburgh, PA, USA 15213
rohitk , cprose @ cs.cmu.edu

Abstract

Instructional efficacy of automated Conversational Agents designed to help small groups of students achieve higher learning outcomes can be improved by the use of social interaction strategies. These strategies help the tutor agent manage the attention of the students while delivering useful instructional content. Two technical challenges involving the use of social interaction strategies include determining the appropriate policy for triggering these strategies and regulating the amount of social behavior performed by the tutor. In this paper, a comparison of six different triggering policies is presented. We find that a triggering policy learnt from human behavior in combination with a filter that keeps the amount of social behavior comparable to that performed by human tutors offers the most effective solution to the these challenges.

1 Introduction

While Conversational Agents have been shown to be an effective technology for delivering instructional content to students in a variety of learning domains and situations (Grasser et. al., 2005; Kumar et. al., 2007; Arnott et. al., 2008), it has been observed that students are more likely to ignore and abuse the tutor in a collaborative learning setting (with 2 or more students) compared to the case of one-on-one tutoring (Bhatt et. al., 2004; Kumar et. al., 2007). In our prior work (Kumar et. al., 2010a), we have addressed this problem by employing agents that are capable of performing both instructional behavior as well as social behavior. In our initial implementation, the social behavior performed by these agents was composed of eleven social interaction strategies that were triggered by a set of hand crafted rules (Kumar and Rosé, 2010b). Section 2 provides additional details about these strategies.

Comparison between the social behavior triggered by our hand crafted rules and that triggered by a human tutor revealed significant perception benefits (more likeable, higher task satisfaction, etc.) for the human triggering policy. Also, the students in a wizard-of-oz condition who interacted with the tutors whose social behaviors were triggered by humans had better learning outcomes (0.93σ) with respect to a No social behavior baseline. The condition where students interacted with the rule-based automated tutors was also significantly better (0.71σ) than the No social behavior baseline in terms of learning outcomes. While the learning outcomes of the rule-based tutors was not significantly worse than the human tutor, in combination with the perception outcomes, we see the potential for further improvement of conversational agents by employing a better triggering policy.

Building on these prior results, in this paper we explore a way to improve the effectiveness of socially capable tutor agents that uses a triggering policy learnt from a corpus of human behavior. The underlying hypothesis of this approach is that a human-like triggering policy would lead to improvements in the agent's performance and percep-

Proceedings of the SIGDIAL 2011: the 12th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 227–238,
Portland, Oregon, June 17-18, 2011.

tion ratings compared to a rule-based triggering policy. As a first step towards verifying this hypothesis, we learnt a collection of triggering policies from a corpus of human behavior. While the focus of this paper is to evaluate the most human-like triggering policy learnt from data in terms of its perception benefits and learning outcomes, Section 4 summarizes our efforts on learning triggering policies.

Before we discuss the details of the evaluation we conducted, Section 3 presents an analysis of mediating factors that provides insights into the reasons behind the effectiveness of social behavior. The design and procedure of the user study we conducted to evaluate the learnt triggering policies is described in Section 5. Finally, Section 6 discusses the results of this evaluation.

2 Social Interaction Strategies

In our prior work (Kumar et. al., 2010; Ai et. al., 2010; Kumar et. al., 2011), we have developed and evaluated automated tutors for two different educational domains equipped with eleven social interaction strategies. These strategies, listed in Table 1, correspond to three positive socio-emotional interaction categories identified by Bales (1950): Showing Solidarity, Showing Tension Release and Agreeing.

Appendix A shows excerpts of an interaction between three students and a tutor during a college freshmen mechanical engineering learning activity. The shaded turns demonstrate realizations of some of the eleven social interaction strategies.

Turns 7-12 shows the tutor initiating and participating in group formation using Strategy 1a (Do Introductions) by greeting the students and asking for their names. In turn 53, the tutor is employing Strategy 3b (Show Comprehension / Approval) in response to a student opinion expressed in turn 52. When one of the students becomes inactive in the interaction, the tutor uses strategy 1e (Encourage) realized as a targeted prompt shown in turn 122 to elicit a response from the inactive student. Turn 148 demonstrates Strategy 1d (Complement / Praise) to appreciate student participation in a conceptual tutoring episode that concluded at turn 147. Finally, turn 152 shows a realization of Strategy 2c (Express Enthusiasm, Elation, Satisfaction) which is tied to either the start or the end of lengthy problem solving steps in the learning activity such as calculating the outcome of certain design choices made by the students during the learning activity.

1. Showing Solidarity *Raises other's status, gives help, reward*
1a. Do Introductions *Introduce and ask names of all participants*
1b. Be Protective & Nurturing *Discourage teasing*
1c. Give Re-assurance *When student is discontent, asking for help*
1d. Complement / Praise *To acknowledge student contributions*
1e. Encourage *When group or members are inactive*
1f. Conclude Socially

2. Showing Tension Release *Jokes, laughs, shows satisfaction*
2a. Expression of feeling better *After periods of tension, work pressure*
2b. Be cheerful
2c. Express Enthusiasm, Elation, Satisfaction *On completing significant steps of the task*

3. Agreeing *Shows passive acceptance, understands, concurs, complies*
3a. Show Attention *To student ideas as encouragement*
3b. Show Comprehension / Approval *To student opinions and orientations*

Table 1. Social Interaction Strategies for three social-emotional interaction categories

3 Effect of Social Behaviors

We have observed the the use of social behavior in collaborative learning settings can help in improving learning outcomes (Kumar et. al., 2010a). This observation is not necessarily consistent with intuition. On the one hand, we could attribute this to the ability to manage the user's (student) attention towards the agent (tutor) by displaying the social behavior. At the same time, we could argue against the use of social behavior by considering it as a potential distraction from the task. In an earlier experiment (Kumar et. al., 2010c), we found that it is important to keep the amount of social behavior at an optimal level in order to observe the learning

benefits of the social behavior and to avoid its distraction costs.

1	Tutor	Are you paying attention to this?
2	S006	of course
3	S001	yes
4	Tutor	What would happen if the stress in our Wrench is too high?
5	S006	fail
6	S001	it'd fail
7	S014	it would break.
8	Tutor	That's right. The wrench can break!
9	Tutor	A wrench that can break in normal use is not safe!
10	S006	:(
11	Tutor	In general, to keep a wrench safe, we want to keep the stress in the wrench high or low?
12	S006	lowwww
13	S001	low
14	Tutor	Right, stress should be kept low.
15	Tutor	Which material property determines when the stress is too high?
16	S001	yield strength?

Table 2. Excerpt of a Conceptual Tutoring Episode

In this section, we present a corpus analysis to discover an underlying model of how the social behavior affects learning outcomes. We used a corpus of collaborative learning interactions between teams of three or more students and a tutor. The interactions involve the teams performing a learning activity which lasts approximately 40 minutes. During this time, the tutors provide directions, hints and conceptual tutoring. Each team undergoes up to four conceptual tutoring episodes. An episode is a system-initiated conversation during which the tutor leads the students through a directed line of reasoning to help them reflect upon a concept related to the learning activity. An excerpt of a tutoring episode discussing the relationship between stress and safety is shown in Table 2.

3.1 Coding Tutoring Episodes

Each turn in all the tutoring episodes of the 32 interactions between a team of students and an automated tutor were annotated using a coding scheme described here. The tutor turns were categorized as either Respondable (TR) if the students were expected to the respond to that tutor turn or Not Respondable (TU) otherwise. In Table 2, all the shaded turns are labeled as Respondable.

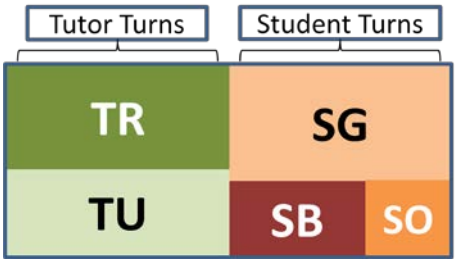

Figure 1. Venn Diagram of Episode Turn Annotations

Student Turns are categorized into one of three categories. Good turns (SG) identifies turns where the students are showing attention to a respondable tutor turn (e.g. Turn 2 & 3 in Table 2) or the students are giving a correct or an incorrect response to a direct question by the tutor (e.g. Turns 5, 6, 7, 12, 13 & 16). Counterproductive (Bad) student turns (SB) include students abusing the tutor or ignoring the tutor (e.g. talking to another student when the students are expected to respond to a tutor turn). Student turns that are not categorized as Good or Bad are labeled as Other (SO). Turn 10 is an example of SO because it is a response to a tutor turn (9) where no student response is expected. Figure 1 shows a Venn diagram of the different annotations. All five categories are mutually exclusive.

3.2 Structural Equation Modeling

In order to discover an underlying model of how the use of social behavior affects student learning, we used a structural equation modeling (SEM) technique (Scheines et. al., 1994).

Data: To measure learning outcomes, our data comprised of scores from pre-test and post-test administered to 88 students who were part of the 32 teams whose data was annotated for this analysis. We normalized the number of Good (SG) and Bad (SB) student turns by the number of Respondable (TR) tutor turns and included normalized SG (*nSG*) and normalized SB (*nSB*) as measures of interaction characteristics of each student in our dataset. Total number of social turns performed by the tutor in each interaction was included as a characteristic of social behavior displayed by the tutor. Finally, the total amount of time (in seconds) that

the students spent on the tutoring episodes was included as a characteristic of the interaction quality during the tutoring episodes.

Prior Knowledge: The only prior knowledge input to the model stated that the pre-test occurs before the post-test.

Discovered Models: We used Tetrad IV to discover a structural equation model in the data comprising of 6 fields (*PreTest*, *PostTest*, *nSG*, *nSB*, *SocialTurns*, *EpisodeDuration*) for each of the 88 students. Figure 2 shows the structural equation model discovered by Tetrad using the dataset described above. p-Value of 0.46 for this model confirms the hypothesis used by Tetrad for its statistical analysis i.e. the model was not discovered randomly. Note that unlike other statistical tests, SEM models built using Tetrad are evaluated as significant if the p-Value is greater than 0.05. The numbers on the arrows are correlation coefficients and the numbers on the boxes indicate mean values for each variable.

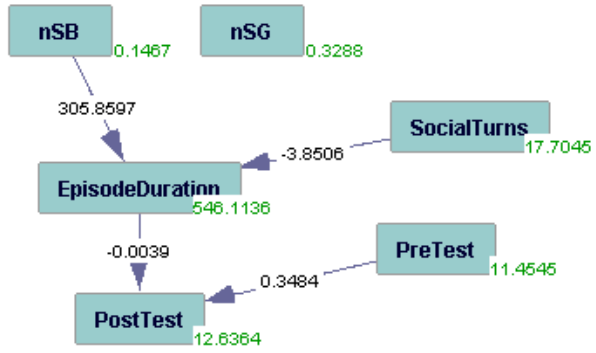

Figure 2. SEM discovered using all 6 variables in our dataset

Besides the obvious causal effect of *PreTest* score on *PostTest* score, we find that as the duration of the tutoring episodes (*EpisodeDuration*) increases, the learning outcomes deteriorate. We notice that an increase in the normalized number of Bad student turns increases *EpisodeDuration* indicating that students who abuse or ignore the tutor are likely to not pay attention to the learning content presented during the tutoring episodes, hence prolonging the tutoring episode as the tutor tries to get the students through the instructional content. Furthermore, we observe that social behavior helps in counteracting the negative learning effect of Bad interaction behaviors of the students. Tutors that perform social behavior are capable of managing the student's attention and get the students through the tutoring episode faster.

3.3 Discussion

The SEM analysis discussed in the previous section helps us better understand the relationship between the use of social behavior and student learning in a collaborative learning setting. Let's consider the duration of the tutoring episodes as an indicator of the students' attention to the tutor (higher duration ⇒ lower attention). We see that social behavior helps in managing the students' attention, which may be affected negatively by counterproductive/bad interaction behavior from the students.

Besides suggesting that social behavior could be a useful strategy for directing student attention, it also suggests that social behavior may not serve this function where counterproductive student behavior is not present or where it does not occur enough to negatively impact task behavior. This is because a minimum amount of time needs to be spent on each tutoring episode to deliver the instructional of the concept being discussed. In the absence of counterproductive student behavior, episode duration may be close to that minimum.

Also, in an earlier analysis (Hua et. al., 2010) in a different learning domain where the social behaviors described in Section 2 were employed, we have observed that the number of abusive/negative comment made by the students about the tutor during the interaction were significantly higher in a condition where the tutors performed a high amount of social behavior. This suggests that the relationship between the *SocialTurns* and *EpisodeDuration* variables may not be linear in extreme cases and emphasizes the importance of performing an optimal amount of social behavior.

4 Triggering Social Behavior

Aside from designing, implementing and regulating the amount of social behavior performed by automated tutors, one of the challenges involved in the appropriate use of social interaction strategies is that of triggering these strategies only at the most appropriate moments during the interaction. Our initial implementation of these strategies (Kumar & Rosé, 2010b) achieved this using a set

of hand crafted rules that used features such as recent student turns, state of the tutoring plan, etc.

Here we will summarize our efforts on building a better triggering policy using a data-driven approach that models the behavior of human tutors at triggering the social interaction strategies listed in Table 1. Using a corpus of 10 interactions between a group of students and partially automated tutors whose social behaviors were triggered by human tutors, we attempt to learn a triggering policy that predicts when the human tutors will trigger a social strategies. Currently, we focus on only learning a triggering policy that determines if a social behavior should be performed. The choice of which behavior is performed when triggered by the policy is still based on the rules used in our earlier implementation as discussed in Section 5.3.

In order to compare the triggers generated by a policy, we use a binary sequence comparison metric called *kKappa* (Neikrasz & Moore, 2010) developed for evaluating discourse segmentation approaches. The metric allows a soft penalty for misplacing a trigger (or a segment boundary) within a window of *k* turns.

We developed a large margin learning algorithm following McDonald et. al. (2005) that iteratively learns the coefficients of a linear function in the feature space that separates turns where human tutors decided to trigger a social behavior from the rest of the turns. Instead of using an instance-based objective function (like square-loss), our algorithm maximizes the *kKappa* metric over a provided training set. The function learnt this way can be used as a triggering policy by using it at every turn during an interaction to predict if a human tutor would trigger a social behavior. We used a collection of automatically extractable features that represent the lexical and semantic content of recent student and tutor turns, current discourse state and activity levels of the students.

While details of the objective evaluation of the various learnt triggering policies is beyond the scope of this paper, we found that the best performing strategy (k-$\kappa = 0.13$) was significantly better than a random baseline (k-$\kappa = 0.01$) as well as the rule based triggering policy (k-$\kappa = -0.09$) used in our initial implementation. Also, the policy learnt by our algorithm outperformed policies learnt by algorithms such as Linear Regression (k-$\kappa = 0.00$) and Logistic Regression (k-$\kappa = 0.05$) that use instance-based loss metrics (Hall et. al., 2009).

5 User Study

Here we will present an experiment we conducted to evaluate the effectiveness of various ways to trigger social behavior discussed in Section 4. This experiment is a step towards verifying the hypothesis that a human-like triggering policy could outperform a rule-based triggering policy that was used in our earlier experiments (Kumar et. al., 2010a). We use the same interactive situation for the experiment presented here as in our earlier work. Freshmen mechanical engineering students enrolled at an American university participate in a computer-aided engineering lab that is divided into three parts, i.e., Computer-Aided Design (CAD), Computer-Aided Analysis (CAA) and Computer-Aided Manufacturing (CAM). Students practice the use of various engineering software packages for all three parts as they design, analyze and manufacture an Aluminum wrench. Our experiment is conducted during the second part (CAA) of the lab.

5.1 Procedure & Materials

The Computer-Aided Analysis lab comprises of two activities. The first activity involves analyzing a wrench design given to the students by specifying certain loading conditions and simulating the stresses and deformations in the wrench. Students are led by a teaching assistant during this activity. They spend approximately 25 minutes performing this activity. At the end of the analysis activity, the students see a simulation of the stress distribution in the body of the wrench.

After the analysis activity, a pre-test is administered. Each student spends 10 minutes working on the pre-test individually. The pre-test comprises of 11 questions, 8 of which are multiple-choice questions and the other 3 are short essay type questions.

The second activity of the CAA lab is a collaborative design activity. During this activity, students work in teams of three. Student in the same team are seated in separate parts of the lab and can only communicate using a text-based chatroom application (Mühlpfordt and Wessner, 2005). The chatroom application also provides a shared workspace in the form of a whiteboard.

After the pre-test, students are given written instructions describing the collaborative design activity. The instructions ask the students to design a better wrench in terms of ease of use, cost of materials and safety compared to the wrench they ana-

lyzed earlier. The students are expected to come up with three new designs in 40 minutes by varying parameters like dimensions and materials of the wrench. The instructions also include various formulae and data that the students might need to use for their designs. Besides course credit, the instructions mention an additional giftcard for the team that comes up with the best design ($10 for each member of the winning team).

Students are asked to log in to their respective team's chatroom. They spend the next 40 minutes working on the collaborative design activity. Besides the three students, the chatroom for each team includes an automated tutor. The tutor guides the students through the first two designs suggesting potential choices for dimension and materials for each design. As the design activity progresses, the tutor initiates four conceptual tutoring episodes to help the students reflect upon underlying mechanical engineering concepts like stress, force, moment, safety, etc., that are relevant to the design activity.

Our experimental manipulation happens during this 40 minute segment. The tutor in each team's chatroom is configured to perform social behavior using different triggering policies as specified by the condition assigned to the team. The conditions are discussed in the next section. Irrespective of the condition, each team receives the 4 conceptual tutoring episodes. Every student performs all the steps of this procedure like all other students.

At the end of the collaborative design activity, a post-test and a survey are administered. Students are asked to spend 15 minutes to first complete the test and then the survey. The post-test is the same test used for pre-test. The survey comprises of 15 items shown in Appendix B. The students are asked to rate each item on a 7-point Likert scale ranging from Strongly Disagree (1) to Strongly Agree (7). The 15 items on the survey include 11 items eliciting perception of the tutor. 9 of the 11 items state positive aspects of the tutor (e.g. …*tutor was friendly.*…). The other 2 items stated negative aspects about the tutor (e.g. …*tutor's responses got in the way.*…). Besides the items about the tutor, 2 items elicited the student's rating about the collaborative design activity. The last 2 items were about the student's satisfaction with their performance on the design task.

In total, both the activities that are part of the CAA lab take approximately 1 hour 40 minutes.

5.2 Experimental Design

The teams participating in the experiment described here were divided into six conditions. These conditions determined the triggering policy and the amount of social behavior performed by the automated tutors. Tutors in the **None** condition did not perform any social behavior. Tutors in the **Rules** condition used the same hand crafted rule-based triggering policy employed in our earlier experiment (Kumar et. al., 2010a). Following the results from another experiment (Kumar & Rosé, 2010c), the automated tutors in the Rules condition performed a moderate amount of social behavior (atmost 20% of all tutor turns). On average, the Rules policy triggered 25 social turns per interaction.

The **RandomLow** and **RandomHigh** conditions used a random triggering policy with a social ratio filter to regulate the amount of social behavior. In both the random conditions, the tutor would trigger social behavior using a random number generator to generate the confidence of triggering a social behavior after every turn (by a student or a tutor). In the RandomLow condition, a behavior would be triggered if the confidence was above 0.91. In the RandomHigh condition, a behavior would be triggered if the confidence was above 0.85. On average, the RandomLow condition had 23 behaviors triggered per interaction. About 37 behaviors were triggered in the RandomHigh condition.

The **LearntLow** and **LearntHigh** conditions used the best triggering policy learnt from a corpus of human triggering of social behavior as discussed in Section 4. The same social ratio filter used in the random conditions was used in these two conditions also. As in the case with RandomLow and RandomHigh, different values of a confidence parameter were used for the LearntLow and LearntHigh conditions to control the number of social behaviors triggered. On average, the LearntLow condition had 22 triggers and the LearntHigh condition had 28 triggers.

5.3 Generating Behaviors

The various triggering policies described above for each of our experimental conditions only determine when a tutor agent will perform a social behavior. In order to perform the social behavior in actual use, the agent must not only determine when

a behavior should be triggered, but also determine which behavior should be performed when a trigger is received. Our implementation of the tutor agent used in this experiment provides a continuous stream of scores for each of the eleven social interaction strategies that the tutor can perform. The scores are computed using handcrafted functions that use the same features used in our rule-based triggering policy (Kumar et. al., 2010b). When a social behavior is triggered, a roulette wheel selection is used to determine the strategy to be performed. The circumference of the wheel assigned to each strategy is proportional to the score of each strategy. If the score of all the strategies is zero, a generic social prompt is performed.

6 Results

126 students enrolled in an introductory mechanical engineering course at an American university participated in the experiment described in this paper. The experiment was conducted on two separate days separated by one week. On each day, four sessions of the Computer-Aided Analysis lab were conducted, and students attended only one assigned session. Session assignment was made based on an alphabetic split. The 126 students were divided into 42 teams. 20 teams participated on the first day of the experiment. They were evenly split into four conditions (None, Rules, RandomHigh & LearntHigh). The remaining 22 teams participated on the second day. Out of these, 5 teams each were assigned to the None and RandomLow condition. 6 teams each were assigned to the Rules and LearntLow conditions.

The rest of this section presents detailed results and analysis of this experiment. To summarize, we found that out of the six evaluated policies only the LearntLow policy that uses a triggering model learnt from human triggering data and generates a moderate amount of social behavior is consistently better than the other policies in terms of both performance as well as perception outcomes. Also, the LearntLow policy is found to be most efficient at delivering the instructional content as indicated by the smallest *EpisodeDuration* in Table 5.

6.1 Learning Outcomes

The learning outcomes analysis presented here shows the advantage of using a triggering policy learnt from a corpus of human triggering behavior along with a filtering technique that regulates the amount of social behavior as shown in Table 3.

We first verified that there was no significant difference between the six conditions on the pre-test scores. As in the case of previous experiments using this learning activity, we saw that the learning activity was pedagogically beneficial to the students irrespective of the condition. There was a significant improvement in test scores between pre-test and post-test { $p < 0.0001$, $F(1,250) = 26.01$, effect-size = 0.58σ }.

There was no significant effect of the condition assigned to each team on the total test scores. However, there was a significant effect on the test scores of short-essay type questions using the pre-test score as a covariate and the condition as a factor { $p < 0.05$, $F(5, 119) = 2.88$ }. The adjusted post test scores for the short essay type questions and their standard deviations are shown in Table 3. Post-hoc analysis showed that the LearntLow condition was significantly better than LearntHigh condition { effect-size = 0.65σ }. Also, RandomLow condition was marginally better than LearntHigh condition { $p < 0.07$, effect-size = 0.62σ }.

	Mean	**St.Dev.**
LearntLow	5.12	0.54
RandomLow	5.06	0.67
None	4.75	1.13
RandomHigh	4.59	1.09
Rules	4.38	0.89
LearntHigh	3.98	1.74

Table 3. Mean and Standard Deviation of Adjusted Post Test Scores for Short Essay Type Questions

This result further supports the observation from our earlier experiment (Kumar & Rosé, 2010c) which demonstrated that importance of performing the right amount of social behavior. Both RandomLow and LearntLow conditions employ the non-linear social ratio filter which keeps the amount of allowed social behavior at a level comparable to the amount of social behavior performed by human tutors.

Since the primary objective of the experiment described here was to evaluate a learnt triggering policy with respect to a rule-based triggering policy, we repeated the ANCOVA for the short essay type question using data from only the Rules,

LearntLow and LearntHigh conditions. We found a significant effect of condition on the post-test score using pre-test score as a covariate { $p = 0.01$, $F(2,62) = 4.98$ }. A post-hoc analysis showed that the LearntLow condition was significantly better than the LearntHigh condition as above and the LearntLow condition was marginally better than the Rules condition { $p \approx 0.08$, effect-size = 0.84σ }. We observe that a triggering policy learnt from human triggering behavior can achieve a marginal improvement on learning outcomes compared to our existing rule-based triggering policy. This is consistent with our hypothesis.

6.2 Perception Ratings

We averaged the student's rating for the 11 items about the tutor into a single tutor rating measure used here. Rating on the two negative statements about the tutor were inverted (7→1, 6→2, and so on) for this calculation.

	Mean	St.Dev.
Rules	4.74	1.45
LearntLow	4.56	1.58
None	4.42	1.49
RandomHigh	3.74	1.63
LearntHigh	3.55	1.26
RandomLow	3.18	0.91

Table 4. Mean and Standard Deviation of Tutor Ratings

We found a significant effect of condition on the tutor ratings { $p < 0.01$, $F(5,120) = 3.83$ }. Table 4 shows the mean and standard deviations of tutor ratings for each condition. Post-hoc analysis showed that only the Rules condition was significantly better than the RandomLow condition. Also, we found that Rules was marginally better than LearntHigh condition { $p < 0.08$ } and both LearntLow and None conditions was marginally better than RandomLow condition { $p < 0.08$ }.

While we did not see a significant improvement in perception due the use of a learnt triggering policy when compared to a rule-based triggering policy, we find an advantage over using a random triggering policy (RandomLow) which was as good as a learnt policy on the learning outcomes. The results from the tutor's perception ratings further support the importance of timing and regulating the amount of social behavior.

We did not find any significant effect of condition on the ratings about the design activity or student's task satisfaction.

6.3 Analysis of Tutoring Episodes

In order to understand the results from the experiment presented in this paper, we applied the structural equation model discussed earlier (Figure 2) to the data collected from our current experiment. Figure 3 shows the model for our current experiment (p=0.4492). Only four variables were used because the annotations of good and bad student behavior are not available at this time.

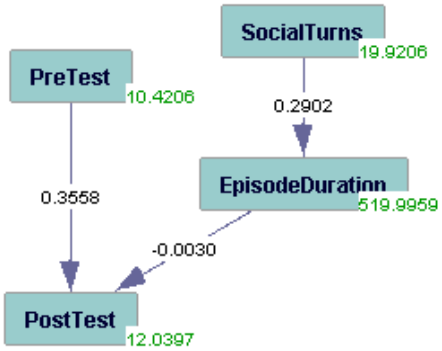

Figure 3. SEM applied to data from this experiment

	Mean	St.Dev.
RandomHigh	540.80	49.50
LearntHigh	534.80	61.00
None	523.88	41.54
Rules	519.80	102.70
RandomLow	519.20	74.40
LearntLow	484.00	69.80

Table 5. Mean and Standard Deviation of Duration of Tutoring Episodes

We see that most of the model parameters (p-Value, means & correlations) are similar to parameters for the model shown in Figure 2. However, the correlation between *SocialTurns* and *EpisodeDuration* is much smaller. Also, note that the mean of *EpisodeDuration* is smaller compared to that in Figure 2 which indicates that lesser counterproductive behavior was displayed by the students in this experiment. The conceptual tutoring episodes are operating closer to the minimum episode duration which leaves a smaller room for improvement by

the use of social interaction strategies. As discussed in Section 3.3, this explains the smaller correlation between *SocialTurns* and *EpisodeDuration* in Figure 3.

Table 5 shows the mean and standard deviations of the duration of tutoring episodes for each condition. Even though the differences are not significant, the LearntLow policy has the lowest duration indicating higher student attention than the other conditions.

7 Discussion

Prior work in the field of human-human interaction and human-machine interaction in the form of dialog systems has emphasized the importance of timing the display of behavior to achieve natural and/or productive interactions. In general, timing of interactive behaviors (verbal as well as non-verbal) has been studied in the context of joint activities being performed by the participants. Behaviors are timed to achieve and maintain coordination between the participants (Clark, 2005). Specifically, among other topics, timing of low-level (signal) interaction like turn-taking has been the subject of several investigations (Raux & Eskenazi, 2008; Takeuchi et. al., 2004).

On the other hand, the use of social behavior by conversational agents to support students has been proposed (Veletsianos et. al., 2009; Gulz et. al., 2010). Work in the area of affective computing and its application to tutorial dialog has focused on identification of student's emotional states and using those to improve choice of behavior performed by tutors (D'Mello et. al., 2005). Our prior work (Kumar et. al., 2010; Kumar et. al., 2007) has shown that social behavior motivated from empirical research in small group communication (Bales, 1950) can help in effectively supporting students in collaborative learning settings. Use of social interaction in other applications of conversational agents besides education has been investigated (Bickmore et. al., 2009; Dybala et. al., 2009; Dohsaka et. al., 2009).

The experiments presented here bridges these two tracks of research specifically proposing a solution to the challenge of timing social behavior in the context of a supporting collaborative learning. Compared to the work on timing signal-level joint activities like turn-taking, this work focuses on the timing of joint activities at the conversation level. The success of our algorithm at learning a model of timing conversational behaviors in the context of an interactive task could potentially offer a general approach for realizing such behaviors in other conversational agents.

8 Conclusion

In this paper, we presented an experiment that compared the effectiveness of several social behavior triggering policies. Specifically, we compared a triggering policy learnt from a corpus of human triggering behavior to a rule-based policy which has previously been shown to be successful at triggering effective social behavior in a collaborative learning activity.

The presented experiment provides further evidence in support of the intuition that timing of social behavior and regulating the amount of social behavior are critical to improving performance and perception outcomes. A triggering policy based on human-like timing in combination with a filter that attempts to keep amount of social behavior at the same level as human tutors was shown to be marginally better than the rule-based policy on learning outcomes. Also, on perception measures, we found that the human-like policy is marginally better than a random triggering policy which uses the same filter to control the amount of social behavior. Only the learned model provides a win both on learning and on perception measures.

In order to better understand the effect of use of social behavior by automated tutors on student's learning outcomes, we presented a structured model which suggests that social behavior helps in achieving higher learning outcomes by allowing the tutor to better manage the student's attention. Following this model, we saw that a human-like triggering policy is able to achieve higher student attention as indicated by the smaller duration of tutoring episodes.

We found a significant negative correlation { coefficient = -0.20, $p < 0.05$ } between the tutor's perception rating and number of social behaviors triggered when none of the social interaction strategies were applicable. As next steps, our best triggering policy could be potentially further refined by achieving a closer integration of the triggering model with the social behavior generation mechanism to prevent triggering when none of the eleven strategies could be generated.

References

Hua Ai, Rohit Kumar, Dong Nguyen, Amrut Nagasunder and Carolyn P. Rosé, 2010, Exploring the Effectiveness of Social Capabilities and Goal Alignment in Computer Supported Collaborative Learning, Intelligent Tutoring Systems, Pittsburgh, PA

Elizabeth Arnott, Peter Hastings and David Allbritton, 2008, Research Methods Tutor: Evaluation of a dialogue-based tutoring system in the classroom, Behavior Research Methods, 40 (3), 694-698

Robert F. Bales, 1950, Interaction process analysis: A method for the study of small groups, Addison-Wesley, Cambridge, MA

Khelan Bhatt, Martha Evens, Shlomo Argamon, 2004, Hedged responses and expressions of affect in human/human and human/computer tutorial interactions, CogSci, Chicago, IL

Timothy Bickmore, Daniel Schulman and Langxuan Yin, 2009, Engagement vs. Deceit: Virtual Humans with Human Autobiographies, Proc. of Intelligent Virtual Agents, Amsterdam, Netherlands

Herbert H. Clark, 2005, Coordinating with each other in a material world, Discourse Studies, 7 (4-5), 507-525

Sidney K. D'Mello, Scotty D. Craig, Barry Gholson, Stan Frankin, Rosalind Picard, Arthur C. Graesser, 2005, Integrating Affect Sensors in an Intelligent Tutoring System, Wksp on Affective Interactions: The Computer in the Affective Loop, IUI, San Diego, CA

Pawel Dybala, Michal Ptaszynski, Rafal Rzepka and Kenji Araki, 2009, Humoroids: Conversational Agents that induce positive emotions with humor, AAMAS, Budapest, Hungary

Kohji Dohsaka, Ryoto Asai, Ryichiro Higashinaka, Yasuhiro Minami and Eisaku Maeda, 2009, Effects of Conversational Agents on Human Communication in Though Evoking Multi-Party dialogues, SIGDial 2009, London, UK

Agneta Gulz, Annika Silvervarg and Björn Sjödén, 2010, Design for off-task interaction - Rethinking pedagogy in technology enhanced learning, Intl. Conf. on Advanced Learning Technologies, Tunisia

Arthur C. Graesser, Patrick Chipman, Brian C. Haynes, and Andrew Olney, 2005, AutoTutor: An Intelligent Tutoring System with Mixed-initiative Dialogue, IEEE Transactions in Education, 48, 612-618

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann and Ian H. Witten, 2009, The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1

Rohit Kumar, Jack L. Beuth and Carolyn P. Rosé, 2011, Conversational Strategies that Support Idea Generation Productivity in Groups, 9th Intl. Conf. on Computer Supported Collaborative Learning, Hong Kong

Rohit Kumar, Hua Ai, Jack Beuth and Carolyn P. Rosé, 2010a, Socially-capable Conversational Tutors can be Effective in Collaborative-Learning situations, Intelligent Tutoring Systems, Pittsburgh, PA

Rohit Kumar and Carolyn P. Rosé, 2010b, Engaging learning groups using Social Interaction Strategies, NAACL-HLT, Los Angeles, CA

Rohit Kumar and Carolyn P. Rosé, 2010c, Conversational Tutors with Rich Interactive Behaviors that support Collaborative Learning, Workshop on Opportunities for intelligent and adaptive behavior in collaborative learning systems, ITS 2010, Pittsburgh, PA

Rohit Kumar, Gahgene Gweon, Mahesh Joshi, Yue Cui and Carolyn Rosé, Supporting students working together on Math with Social Dialogue, Workshop on Speech and Language Technology in Education, Farmington, PA, 2007

Ryan McDonald, Koby Crammer and Fernando Pereira, 2005a, Online large-margin training of dependency parsers, Proc. of ACL, 91-98, Ann Arbor, MI

Martin Mühlpfordt and Martin Wessner, 2005, Explicit referencing in chat supports collaborative learning, Proc. Computer Support for Collaborative Learning (CSCL)

John Niekrasz and Johanna D. Moore, 2010, Unbiased Discourse Segmentation Evaluation, Proc. of SLT, Berkeley, CA

Antoine Raux and Maxine Eskenazi, 2008, Optimizing Endpointing Thresholds using Dialogue Features in a Spoken Dialogue System, SIGDial 2008, Columbus, OH

Richard Scheines, Peter Spirtes, Clark Glymour, and Christopher Meek, 1994, TETRAD II: Tools for Discovery, Lawrence Erlbaum Associates, Hillsdale, NJ

Masashi Takeuchi, Norihide Kitaoka and Seiichi NakagawaM, 2004, Timing detection for realtime dialog systems using prosodic and linguistic information, Intl Conf. on Speech Prosody 2004, Nara, Japan

Tetrad IV, http://www.phil.cmu.edu/projects/tetrad/tetrad4.html

George Veletsianos, Charles Miller and Aaron Doering, Veletsianos, 2009, EnALI: A Research and Design Framework for Virtual Characters and Pedagogical Agents, Journal of Educational Computing Research, 41(2), 171-194

Appendix A. Excerpts of an interaction between three students and an automated tutor demonstrating some of the social interaction strategies listed in Table 1.

7	Tutor	Hi, I am your tutor for today's lab.
8	Tutor	Lets introduce ourselves briefly. I am Avis.
9	S083	Hey Avis! I'm `_name1_`
10	S073	Im `_name2_`
11	S089	i'm `_name3_`
12	Tutor	Its nice to meet you all. :)
13	Tutor	Let's get started by reviewing the base design in your worksheet.

Strategy 1a: Do Introductions

48	S083	R we using the same material?
49	S073	I assume so
50	S073	just changing the length
51	S089	yeah we have multiple design steps
52	S089	so probably for now
53	Tutor	cool :)
54	S083	O ok

Strategy 3b: Show Comprehension / Approval

119	Tutor	Is this a safe wrench?
120	S073	and then is the same for design 2
121	S073	so yes the wrench is safe
122	Tutor	`_name1_` ... any thoughts you'd like to contribute?
123	S083	its a safer wrench if its in steel

Strategy 1e: Encourage (inactive members)

143	S073	high
144	Tutor	Right, higher yield stress is better.
145	S089	so steel or titanium
146	S089	but have fun paying for that...
147	Tutor	All other things being equal, you want to choose a stronger material.
148	Tutor	It's good to have your contributions to the discussion :-)
149	S073	yay

Strategy 1d: Complement / Praise

150	S073	:)
151	Tutor	Let's improve design 1 by using Steel for our 2nd `_truncated_`
152	Tutor	I am happy to be working with our team
153	S083	thanks :-)
154	Tutor	You can start calculating the fields in the worksheet `_truncated_`
155	S089	woo...

Strategy 2c: Express Enthusiasm, Elation, Satisfaction

Appendix B. Survey administered to the participants at the end of the Collaborative Design Activity

Using the following scale, Indicate to what extent you agree with each of the following items.

1 Strongly Disagree	2 Mostly Disagree	3 Somewhat Disagree	4 Neutral	5 Somewhat Agree	6 Mostly Agree	7 Strongly Agree

The tutor was part of my team.	1	2	3	4	5	6	7
The tutor provided good ideas for the discussion.	1	2	3	4	5	6	7
The tutor received my contributions positively.	1	2	3	4	5	6	7
The tutor was friendly during the discussion.	1	2	3	4	5	6	7
The tutor responded to my contributions.	1	2	3	4	5	6	7
The tutor helped in lowering the tension in my group.	1	2	3	4	5	6	7
The tutor was paying attention to our conversation.	1	2	3	4	5	6	7
Overall, I liked the tutor very much.	1	2	3	4	5	6	7
I think the tutor was as good as a human tutor.	1	2	3	4	5	6	7
I often ignored what the tutor was saying.	1	2	3	4	5	6	7
The tutor's responses got in the way of our conversation.	1	2	3	4	5	6	7
The design challenge was exciting.	1	2	3	4	5	6	7
I did my best to come up with good designs.	1	2	3	4	5	6	7
I am happy with the discussion I had with my group.	1	2	3	4	5	6	7
Overall, we were successful at meeting our goals during the design challenge.	1	2	3	4	5	6	7