Choosing Optimal Levels of Social Interaction - Towards creating Human-like Conversational Tutors

Rohit Kumar, Hua Ai, Carolyn P. Rosé

Language Technologies Institute, Carnegie Mellon University,
5000 Forbes Avenue, Pittsburgh, Pennsylvania 15213
{ rohitk , huaai , cprose } @ cs.cmu.edu

Abstract. Recent results show that socially-capable conversational tutors can be effective support in collaborative learning situations. However, in comparison to human-level social capability, we find scope for improving the tutors both in connection with learning (performance) as well as liking (perception). In this paper, we describe work towards improving state-of-the-art conversational tutors by shrinking the gap between automated and human-level social capabilities. An experiment that explores the effect of amount of social behavior is described. Besides confirming the learning effects of socially-capable tutors from a previous study, we report on the optimal levels of social and task interaction to improve learner performance.

Keywords: social interaction, conversational agents, collaborative learning

1 Introduction

Conversational Tutors used in state of the tutorial dialog system have been shown to be effective support for individual learners [1] [2] as well as collaborative learning groups [3]. Investigations in this promising educational technology have largely focused on appropriate delivery of instructional content relevant for the learning task. However, research in the field of small group communication has shown that group members participate in both task-related as well as socio-emotional interaction.

In recent work [4], we have shown that tutors capable of performing social interaction while working with groups can be significantly better than tutors that have no social capability. Specifically, we found that a tutors augmented with human social capability were able to achieve a 0.93 standard deviations (σ) effect relative to non-social tutors. An automated implementation [4] of eleven social interaction strategies (listed in Table 1) motivated from three positive socio-emotional interaction functions observed in small group interactions [5] using the Basilica architecture [6] achieved a relative effect of 0.71σ compared to the same baseline. Even though the effect of the human level social capability was higher than the automated social capability relative to the baseline, the difference between them was not statistically significant. However, on the perception metrics, we noticed that the human capability outperformed our implementation of the social-capable automated tutor on most measures.

Table 1. Social Interaction Strategies based on three of Bales' Positive Socio-Emotional Interaction Categories

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 task steps*

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*

Both the poor performance on the perception metrics and the lower effect size suggest scope for improving the automated implementation of the social capability we have built in the tutors. Towards this goal, we adopt an anthropomorphic approach and attempt to make our automated socially-capable tutors resemble the human tutors in behavior. Besides the evidence of better performance of the human social capability [4], we adopted this approach because of its practical applicability given the availability of human social behavior data in similar conversational settings.

One of the observations we made by comparing the social behavior actually displayed by the human and the automated tutors was that the human tutors displayed significantly more instances of social turns compared the automated tutor. In section 2 of this paper, we will present more details of this analysis and describe the implementation of a new automated tutor that tries to match the human social behavior in terms of the quantity of social turns. In section 3, we will describe an experiment we conducted to evaluate this new tutor. Section 4 reports on the results of this experiment both on learning performance as well as perception scales. We will also discuss the optimal level of social behavior a tutor should perform in a typical collaborative learning conversational situation we have used for our experiment. This will be followed by conclusions and remarks towards next steps in the anthropomorphic development of socially capable conversational tutors.

2 Human-level Social Behavior

From a controlled experiment [4] we conducted to compare tutors with different levels of social capabilities, we found that *Human* tutors contributed significantly more social interaction turns to the interaction with the students compared to our socially capable tutors that implement the strategies listed in Table 1. Table 2 compares the automated and the Human tutors w.r.t display of social turns. First of all, we note that the human tutors perform significantly more social turns related to showing solidarity (strategies 1a - 1f) and agreeing (strategies 3a-3c). Also, tension release strategies (2a-2c) were used quite infrequently by the human tutors as opposed to the automated tutors.

Table 2. Average number of social behavior turns displayed by tutor

Behavior	Strategy	Automated	Human
Showing Solidarity	1a-1f	9.25	13.70
Showing Tension Release	2a-2c	5.81	1.77
Agreeing	3a-3b	1.78	4.90

The first observation led us to hypothesize that one of the possible reasons for the shortcomings of the automated tutors on the performance and perception metrics could be attributed to insufficient display of social behavior by those tutors. Table 3 shows an excerpt from an interaction between a team of students and a Human tutor. Names are changed to preserve anonymity. A comparable example of a similar interaction is shown in Table 4 where a team of students is interacting with an automated tutor. In both cases, one of the members of the team responds (turns 65 and 82 respectively) to the tutor checking about the student's progress. The human tutor uses this event as an opportunity to encourage (Strategy 1e) the other group members to contribute (turn 66, 67) unlike the automated tutor.

Table 3. Excerpt of an interaction between a team of students and a Human tutor

	Speaker	Contribution
63	Tutor	Did you manage to calculate the moment already?
65	S66	600?
66	Tutor	talk to your teammates about the design. Discussions are helpful!
67	Tutor	does anyone want to respond to Timmy?
70	S73	yes it is. x2
71	S73	everything is x2
72	Tutor	Thank you Bobby
79	S73	you are welcome avis

Table 4. Excerpt of an interaction between a team of students and an automated tutor

	Speaker	Contribution
80	Tutor	Now let's calculate the maximum stress in the Wrench.
81	Tutor	NOTE: Stress increases proportionately with the Moment.
82	S99	So just 8750x2?
83	Tutor	Since the moment doubled, the maximum stress in design 1 should be doubled too, i.e., 12800lb/in^2

The experiment described in the section 3 attempts to verify (or invalidate) our hypothesis before we consider other possible improvements that could be made to the automated tutors in order to make them more human-like. The effect of the amount of social behavior displayed on outcomes is our first choice for an experiment as it is one of the easiest dimensions to manipulate given our existing implementation of a socially-capable tutor. Further, if this hypothesis was shown to be valid, it has the potential to cover the most ground towards matching human social capabilities. Next, we describe how our automated tutors were modified to vary the amount of social behavior they displayed.

2.1 Implementation of Socially-capable Tutors

Our socially capable tutors are implemented using the Basilica architecture [6] which provides the flexibility to build conversational agents by incrementally adding loosely coupled behavioral components. In the specific case of the tutors used in this work, the architecture has allowed us to build two behavior controllers (f_{plan}, f_{social}) unlike a single controller used in most dialog system architectures. The plan controller (f_{plan}) is similar to the planners used in typical dialogue agent which executes a sequence of steps in order to complete the task (delivering instructions and lessons in this case). The social controller implements the social interaction strategies listed in Table 1. The two controllers coordinate among themselves by blocking each other while performing their functions. More details of this implementation are described in [4]. In the tutors used in the experiment presented here, we did not use the tension release strategies (2a, 2b, 2c) because of their infrequent use by the human tutors.

The amount of social behavior generated by the social controller is regulated using a parameter (social ratio) that specifies the percentage of all turns contributed by the tutor that can be generated by the social controller. In the case of our first implementation of socially capable tutors, this level was set at 20%, i.e., for every 100 turns the tutor displays at most 20 could be generated by the social interaction strategies. In the experiment described in the next section, we use two versions of this tutor with different values of this parameter. The tutor that generates lower amounts of social behavior (*Low*) is configured at 15% social ratio. The other version (*High*) is configured at 30% social ratio which is comparable to the percentage of social turns displayed by the human tutors in our earlier experiment.

3 Experiment

We conducted an experiment to evaluate the effect of amount of social behavior displayed by the automated tutors on performance and perception metrics. The experiment was part of a sophomore Thermodynamics lab project. 106 students enrolled in a sophomore Mechanical engineering course participated in the experiment. The students worked in teams of two to design a Rankine cycle. The experiment was conducted over 3 consecutive days of the same week. Two sessions were held each day. So, different students participated in the six different sessions. With in each session, students were randomly assigned to groups and conditions.

3.1 Procedure and Materials

The procedure for the lab was divided into eight phases. i.) Students were led through a tutorial on using a thermodynamics cycle simulator called CyclePad [7]. ii.) Students read through written material on the subject of Rankine Cycle and green engineering. iii.) Students used the CyclePad software to analyze the response of the cycle in terms of its efficiency, net power, waste hear and steam quality with respect to various system properties like temperature and pressure. During this phase, students followed along with our lab coordinator. iv.) Following the tutorial, students filled out a motivation questionnaire (5 items) and a pre-test (30 items). v.) Students were given a tutorial on a collaboration software called ConcertChat [8] which they used in the next phase. vi.) Next, the students logged into private ConcertChat rooms of their respective teams and started interacting with their teammate and an automated tutor. The students were asked to design a new Rankine cycle by choosing a set of values for the system properties in order to find an optimal output on the response variables. They were told that teams with the best designs will receive gift cards worth $20 as an additional incentive besides class credit which all participants received. To guide their design and to enable systematic interaction with the tutors, the students were asked to follow a worksheet which was designed to guide the students through every system property while considering its effect on each of the responses. vii.) After the collaborative design phase, a post-test (29 items) was administered. They students also responded to a survey designed to elicit student's ratings about the tutor and the design task (among other things) on a 7-point Likert scale. The items used on the survey were similar to those used in [4]. viii.) Finally, the students implemented the designs they came up with during the design phase individually using CyclePad. They were allowed to make further modifications to the design based on the observed responses from the simulator.

3.2 Experimental Design

Our experimental manipulation was part of a larger experiment with multiple independent variables. The manipulation we are concerned with here is with regards to the amount of social behavior (social ratio) the tutors employed in phase (iv) were allowed to display. The student teams (dyads) were randomly assigned to one of three conditions i.e. *None* (0%), *Low* (15%) and *High* (30%). The corresponding values of social ratio for each of these conditions are shown in parenthesis. Conditions were evenly distributed among the teams across sessions. Each team spent the same amount of time on the collaborative design activity (35 minutes).

4 Results & Other Analysis

The automated tutors used in all the three conditions of our experiment are compared with respect to student learning outcomes, their ratings as reported by the students on the survey and the differences in student's rating about the collaborative design task.

4.1 Learning Outcomes

The pre-test had one additional question than the post-test which was added to make the pre-test and post-test slightly different. This question was not used for calculating pre-test scores. Also, one of the questions on the tests was not used in calculating the test scores as it was very open-ended. Among the remaining 28 questions, 22 were objective (multiple choice questions) and 6 were subjective (brief explanation questions).

An ANOVA using the condition as an independent variable showed the there was no significant difference between the conditions on the total pre-test scores. This was also the case for the scores on the subjective questions and the objective questions. There was a significant improvement in all test scores (total, subjective and objective) between the pre-test and the post-test in all conditions, which shows that in general, the collaborative design activity was beneficial to all students. With respect to the pre-test, the relative effect sizes were 0.79 standard deviations (σ) for the total score, 0.69σ for the objective scores and 0.73σ for the subjective scores. All scores for both the pre and the post tests are shown below in Table 5.

Table 5. Average Pre & Post test scores for each condition
(Standard deviation in paranthesis)

Condition		Pre-Test			Post-Test	
	Total	Objective	Subjective	Total	Objective	Subjective
None	13.94	11.28	2.67	17.72	13.33	4.39
(0%)	*(4.53)*	*(2.91)*	*(2.23)*	*(4.09)*	*(2.47)*	*(2.04)*
Low	14.00	11.38	2.62	18.59	14.77	3.82
(15%)	*(6.15)*	*(4.16)*	*(2.54)*	*(4.72)*	*(3.43)*	*(1.74)*
High	14.08	12.03	2.06	17.72	13.75	3.97
(30%)	*(4.46)*	*(3.13)*	*(1.88)*	*(3.77)*	*(3.07)*	*(1.72)*

Three different ANCOVA models for the three types of scores using corresponding pre-test scores as a covariate and condition and session as independent variables showed that there were no significant differences between the three conditions (*None*, *Low* and *High*) on the total as well as the subjective scores. However, there was a significant effect of the condition variable on the objective scores $F(2, 97)=3.48$, $p < 0.05$. A pairwise Tukey test post-hoc analysis showed that the *Low* (15%) social ratio condition was marginally ($p < 0.07$) better than both *None* (effect size = 0.69σ) and *High* (effect size = 0.55σ) social ratio conditions. The difference between the *None* and the *High* conditions was not significant.

We find a similar effect on one of the learning performance metrics as reported in our previous experiment [4] by an automated tutor with a comparable social ratio (20%). The hypothesis that performance gap between human and automated social tutors can be bridged by performing more social behavior like the human tutors does not hold in the case of learning metrics. Further, we think that the lack of significant differences on the subjective questions is because the tests were very long and the students might have focused more on the objective questions to complete most of the test. This is reflected in the relatively high scores on the objective questions (mean = 13.93) compared to a maximum of 22. In the case of the subjective questions (mean = 4.07), the maximum possible score was 11.

4.2 Ratings about the Tutor and the Learning Task

The survey used to elicit ratings from the students was similar to the survey used in our previous work [4]. However, the survey item about tension release was not used because we did not use the tension release strategies in this experiment (as mentioned in Section 2.1).

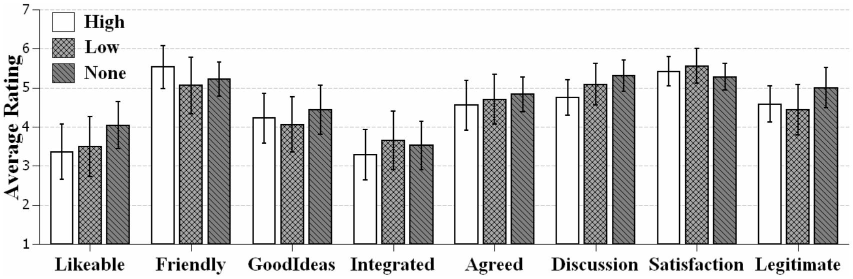

Fig. 1. Average ratings for the Tutor and the Learning Task
95% confidence intervals are also shown

Figure 1 above shows the average rating for the three types of tutors used in our manipulation. The first five items (Likeable, Friendly, Provided Good Ideas, Integrated with Team, Agreed) were related to perception of the tutor by the students. The last three items (Quality of Discussion, Task Satisfaction, Legitimacy of the Task) were about the learning task. None of the differences between the three types of tutors were statistically significant for these perception measures. Once again, we note that the hypothesis that suggests performing higher amount of social behavior to create human-like tutors does not hold for these measures.

4.3 Exposure Effect with Tutors

An additional analysis we were able to perform with the data available from this study was the effect of multiple exposures to automated tutors. Since our studies with engineering students span multiple years and classes, we were able to determine that 27 of our 106 participants had participated in a pilot study in a previous semester. The pilot study employed interaction with automated tutors (with no social capabilities) to teach the students about freshmen mechanical engineering concepts like relationships between forces, moments and stress. By including prior exposure as a binary (yes, no) pseudo-independent variable in the ANCOVA used to model learning outcomes on the objective questions (as described in Section 4.1), we found a significant interaction between the condition and the prior exposure variables F(2, 94) = 3.68, $p < 0.05$. Figure 2 below shows the interaction plot for the two variables.

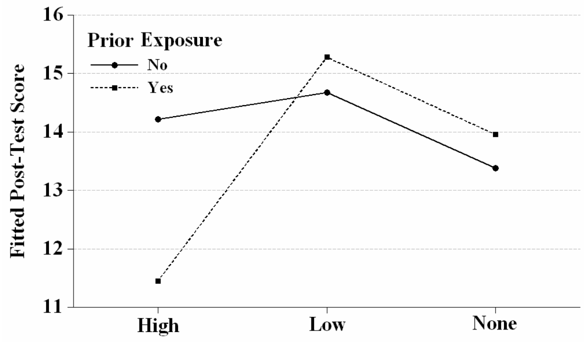

Fig. 2. Interaction between our Experimental manipulation and Prior Exposure to Tutors

We note that tutors that display high amounts of social behavior lead to significantly poor performance for students who have had prior exposure to automated tutors. Relative to the students who do not have prior exposure to such tutors, the effect size is 1σ.

4.4 Estimating the Optimal Amount of Social Behavior

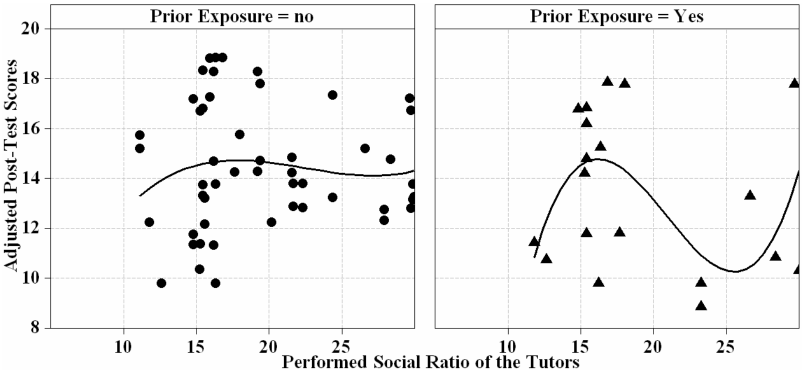

Fig. 3. Scatter plot between Adjusted Post-Test scores and Social Ratio of the tutors in *High* and *Low* conditions

Up until here, we find in general that *high* (30%) social ratio tutors are not significantly different than tutors with no social capabilities (*None*). Also, in the case of students with prior exposure to automated tutors, these (*High*) tutors were significantly worse.

So, why do the *High* tutors lead to poor learning? We found that there was a significant effect of condition on the number of tutorial dialog turns the tutor performed F(2, 98) = 5.01, p < 0.01. A pairwise Tukey test post-hoc analysis showed that in the *High* condition (Mean=76.56, s.d.=9.03) the tutor performed significantly fewer dialog turns compared to the *None* condition (Mean=82.42, s.d.=4.67). The dialog turns performed by the tutors in Low condition was not significantly different from either *High* or *None* conditions (Mean=80.59, s.d.=11.59). Students in the *High* condition didn't move as efficiently through the material and therefore didn't receive all of the instruction. Fewer dialog turns led to lower coverage of domain relevant material during the learning activity, which in turn led to poor performance on the tests.

The above observations suggest the relationship between learning performance and the amount of social behavior displayed by the tutor is non-monotonic. Figure 3 shows cubic polynomial regressions between the adjusted post-test scores and the percentage of social turns performed by the corresponding tutors for each student. Students with and without prior exposure to automated tutors are shown separately. We see that both in the case of students with or without prior exposure to automated tutors, a maxima in performance can be found around 16% performed social ratio.

5 Conclusions

To summarize, we find that the tutors with low social ratio (15%) perform better than the high social ratio (30%) tutors and tutors with no social capabilities on learning outcomes. On perception metrics, these tutors are not significantly different from each other. Both these observations invalidate the hypothesis that matching the display of social behavior with human tutors in quantity will lead to human-like outcomes. Further, the learning result about the *Low* tutors is consistent with our earlier results about socially capable tutors with comparable social ratio.

The poor performance of *High* social ratio tutors suggests that the right amount of social interaction benefits the learning activity by keeping the group's instrumental and expressive needs fulfilled, excessive social interaction becomes a distraction and hinders the task-related interaction (dialogs about lessons in this case). We report empirical values for optimal amounts of social behavior suitable for automated tutors in collaborative learning situations.

Having shown that automated tutors cannot match performance of human tutors merely by matching the amount of social behavior displayed by the human tutors, we turn our investigation to other aspects of human social behavior display. Among the many options as next steps in improving the social capabilities of tutors, we think closer attention needs to be paid to circumstances under which human tutors choose to employ various social strategies and how the display of these strategies is intertwined with task based interaction. For example, in the excerpt shown in Table 3, the decision to elicit participation from other students may be relevant only if Timmy's contribution to recent discussion out weighed contributions of the other students. Another aspect that can be potentially useful in modeling good social behavior by tutors is the study of student responses (or lack of responses) in the data we have collected from recent studies. In Table 3, turn 79 suggests that the tutor's social behavior in turn 72 (Thanking Bobby) was appropriate.

References

1. Graesser, A.C., Chipman, P., Haynes, B. C., Olney, A.: AutoTutor: An Intelligent Tutoring System with Mixed-initiative Dialogue. IEEE Transactions in Education, vol. 48, pp. 612-618 (2005)
2. Arnott, E., Hastings, P., Allbritton, D.: Research Methods Tutor: Evaluation of a dialogue-based tutoring system in the classroom, Behavior Research Methods, vol. 40 (3), pp. 694-698 (2008)
3. Kumar, R., Rosé, C. P., Wang, Y. C., Joshi, M., Robinson, A.: Tutorial Dialogue as Adaptive Collaborative Learning Support. In: Proc. of AI in Education (2007)
4. Kumar, R., Ai, H., Beuth, J. L., Rosé, C. P.: Socially-capable Conversational Tutors can be Effective in Collaborative-Learning situations. Submitted to Intl. Conf. on Intelligent Tutoring Systems, Pittsburgh, PA (2010)
5. Bales, R.F.: Interaction process analysis: A method for the study of small groups, Addison-Wesley, Cambridge, MA (1950)
6. Kumar, R., Rosé, C. P.: Basilica: An architecture for building conversational agents. In: Proc. of NAACL-HLT, Boulder, CO (2009)
7. Forbus, K. D., Whalley, P. B., Evrett, J. O., Ureel, L., Brokowski, M., Baher, J., Kuehne, S. E.: CyclePad: An Articulate Virtual Laboratory for Engineering Thermodynamics. Artificial Intelligence. vol. 114 (1-2), pp. 297-347 (1999)
8. ConcertChat, http://www.ipsi.fraunhofer.de/concert/index_en.shtml?projects/chat