Accuracy of Tutors' Assessments of their Students by Tutoring Context

Stephanie A. Siler (siler@pitt.edu) and Kurt VanLehn (vanlehn@pitt.edu)
Learning, Research and Development Center
University of Pittsburgh
Pittsburgh, PA 51260

Abstract

In this study we compared the accuracy of tutors' assessments of their students' general competence, conceptual knowledge and affective state in two different tutoring contexts: face-to-face (FTF) and computer-mediated (CM). We found that the accuracy of tutors' assessments of their students was dependent on both the type of student information that was assessed, and, to a lesser extent, the tutoring context. Only tutors' assessments of their students' general competence, as opposed to their assessments of their students' individual conceptual knowledge or their students' motivation, was affected by the manipulations.

Introduction

One-to-one human tutoring is generally more effective than classroom instruction (e.g., Bloom, 1984; Cohen, Kulik & Kulik, 1982). Many have assumed that this is in part because tutors can understand their students' domain competence and attitudes and, through this understanding, adapt their instruction to their students. However, some (e.g., Chi, 1996) have questioned this assumption. As part of a larger study testing this adaptive hypothesis, we sought a way to vary tutors' understanding of their students while minimizing disruption to the tutoring situation.

In this study, we varied the amount of experience tutors had with a particular student. In some conditions, each tutor tutored the same student for the entirety of the tutoring session (Same conditions); in other conditions, each tutor tutored four different students (one at a time) in the tutoring session (Different conditions). Thus, tutors in the Same conditions had more experience with an individual student than tutors in the Different conditions. As the results will show, this manipulation did make a difference in the accuracy of some types of tutors' assessments of their students, demonstrating that tutors can assess their individual students during tutoring. A second manipulation was tutoring context. We compared tutoring in a spoken face-to-face (FTF) context with tutoring in a computer-mediated (CM) context, in which tutors and students could not see or hear each other but communicated through typed messages, to determine the effects of the tutoring context on the tutors' assessments of their students' domain competence, conceptual knowledge, and affective states.

Earlier work comparing CM to FTF communication contrasts the amount and content of message production (e.g., Lebie, Rhoades & McGrath, 1996; Ruberg, Moore & Taylor, 1996) and efficiency of the message (e.g., Hausmann & Chi, 2002); however, no work that we are aware of has been done comparing the accuracy of tutors' assessments of their students as a function of tutoring context.

In FTF communication, more sources of information are available to tutors that they may use to assess their students. For example, in a FTF context, prosodic information (e.g., vocal pitch, loudness, turn duration, speaking rate) is available that is not available in a text-only CM context (Litman, 2002). Other types of potentially useful information available only in the FTF context include facial expressions (e.g., puzzled, upset, disinterested), body language (e.g., leaning direction), and nonlinguistic verbalizations (Fox, 1993). All of these may be particularly useful sources of information about the student's affective state. However, there are other sources of information available in a FTF context that may not be useful (e.g., the student's general appearance), or may even impair the tutor's ability to accurately assess the student.

Although tutors in a FTF context have more information available to them, there are some aspects of CM tutoring that may benefit the development of tutors' assessments of their students. For example, in CM tutoring, there is a record of the dialog, allowing tutors to re-read portions of the current dialog they may have initially missed or misunderstood. In a FTF condition, if tutors do not hear or understand messages from their students when they are spoken, this information may be lost to the tutor. Additionally, tutors in CM tutoring may refer back to dialog previously read. Because of repeated exposure to this information, the likelihood of retaining that information increases. In FTF tutoring, tutors must rely on their memories of the past discussion or on notes taken (Lebie et al., 1996). Additionally, in CM tutoring, there is more time between conversational turns (Clark & Brennan, 1996), allowing more opportunity for tutors to think about and more deeply process information.

Our research addressed the question of whether tutors develop more accurate assessments of their students' general domain competence, conceptual knowledge, and affective state in a CM context, a FTF context, or whether

context makes no difference. Having more accurate assessments of their students may lead to tutoring that is more adaptive to the individual students, and thus to more effective tutoring. For example, having accurate assessments of students' general competence may lead to more adaptive tutoring if these assessments influence tutors' choices of the appropriate level of difficulty for the question to ask their student. Having more accurate assessments of students' conceptual knowledge may lead to more adaptive tutoring if tutors use their assessments, for example, when deciding on which concepts to discuss with their students (i.e., those concepts the student does not understand). Finally, being more sensitive to their students' affective states may benefit tutoring provided tutors can maintain positive affect in their students. For example, tutors who realize their student is feeling under-confident may choose to ask their students easier questions than tutors who are not sensitive to their students' level of confidence.[1]

This question also has important implications for natural-language tutors, which primarily have text rather than spoken input from students (Clark, Bratt, Lemon, Peters, Pon-Barry, Thomsen-Gray & Treeratpituk, 2002).

Methods

Participants

Eighty undergraduate and graduate engineering or science majors served as tutors in this study. Eighty undergraduate students who had taken physics in either high school or college served as students. Participants either received course credit for their participation or payment (tutors received $10/hour and students received $7/hour).

Materials

Physics pretest: Definition questions and short-answer questions designed to assess students' knowledge of the concepts relevant to the questions discussed in the tutoring session.

Motivation questionnaire: A revised version of the Motivation for Reading Questionnaire (Wigfield & Guthrie, 1997), adapted to physics. The motivation questionnaire consisted of 21 statements that assessed different dimensions of students' motivation for physics, including their interest in physics, their desire for challenging physics problems, and their confidence in their physics ability. Following each statement was a 7-point scale for students to rate their level of agreement with the statement (1 – "strongly disagree" to 7 – "strongly agree").

Physics Midtest: 23 short-answer questions that assessed students' knowledge of concepts that were relevant to the solutions to the questions discussed in tutoring.

[1] Whether tutors' moves *are* adaptive to their students is one question we are currently investigating.

Procedure

All students were first given the physics pretest. About one week later, they returned for the tutoring session. Half of the 80 tutors discussed 6 physics questions with one student, 20 in a CM context and 20 in a FTF context (Same conditions). The other half of the tutors discussed the same physics questions with four different students, again, 20 in both the CM and FTF contexts (Different conditions). In the Different conditions, the order of students tutored by a given tutor was chosen to maximize differences in total pretest scores and individual conceptual knowledge for consecutive students[2].

In the CM context, each tutor and student communicated using NetMeeting, which has a chat window similar to other common instant messaging programs. Turn-taking was not constrained; the student and tutor could send a message at any time. Netmeeting also has a drawing window in which both tutor and student could draw; anything drawn in the drawing window was visible simultaneously to both. Tutors and students in the FTF context were given paper and pens and were permitted to draw. In both contexts and for all four conditions, tutors were given time limits of 15 minutes to discuss the first 2 questions (segment 1), half an hour to discuss the next two questions (segment 2), 15 minutes to discuss the next question (segment 3), but no time limit to discuss the last (and most difficult) question (segment 4). They were permitted to end discussion before the time limit if they believed the question(s) had been adequately addressed.

Before the last question, students first completed the motivation questionnaire on which they indicated their agreement with each statement on a 7-point Likert scale. Then each student took the physics midtest. As the students worked on the motivation questionnaire and midtest, in another room, each tutor was first instructed to guess his or her student's response to each statement on the motivation questionnaire. Then they were instructed to rank their student's general competence on a 7-point Likert scale (from 1 "well below average" to 7 "well above average"). Finally, to test tutors' assessments of their students' individual conceptual knowledge, each tutor was given a subset of short-answer questions on the midtest and the correct solutions to those questions, broken down by concept. Tutors were instructed to indicate whether they thought the student would demonstrate knowledge of each concept or not[3]. They

[2] If tutors in the Different conditions based their estimates of their future students' competence or conceptual knowledge on their prior student(s), then ordering students so that the last student's competence or conceptual knowledge was maximally different from the penultimate student should reduce the accuracy of the Different tutors' assessments of their final students and magnify evidence that tutors are gaining knowledge of their students in the Same conditions.

[3] Tutors' assessments of students' incorrect beliefs and misconceptions were not investigated because prior research (e.g., Chi, Siler, & Jeong,

could also respond that they were "50/50" (or thought the student was as likely to demonstrate knowledge of the concept as not). Tutors in the Different conditions were given identical instructions with the exception that they were instructed to guess the motivation questionnaire responses, the general competence, and the conceptual knowledge for the student they were about to tutor in the last tutoring segment (but had not tutored before). Tutors' responses were later compared to students' actual responses.

Thus, tutors' knowledge of their students' physics competence, their students' knowledge of individual concepts, and their students' motivation were assessed. Because the motivation questionnaire assessed several dimensions of students' motivation (including students' confidence, interest in physics, and desire for challenging physics problems), tutors' knowledge of various dimensions of their students' motivation was also assessed.

Results

On average, the total tutoring time of the first three segments in the FTF Same condition was significantly less than the total tutoring time of the first three segments in the CM Same condition (Table 1), $t(38) = 5.03$, $p < .001$. Thus, tutors in the CM Same condition spent more time with their student before their knowledge of their student was assessed. However, in the three segments prior to assessing tutors' knowledge of their students, the estimated number of words spoken by students in the FTF Same condition was about 4.5 times the estimated number of words typed in the CM Same condition.[4,5]

Table 1. Mean time of segments 1 – 3 per condition.

Condition	mean time in minutes (SD)
CM Same	**57.53 (6.65)**
FTF Same	**40.75 (13.34)**

When possible, the CM Same and FTF Same conditions were compared with their corresponding Different conditions to assess how much information tutors gained about their students.

(1) *Tutors' assessments of students' relative physics competences: tutor rankings.* In the CM Same condition (but not the CM Different condition), tutors' rankings of their student's physics competence was positively correlated with students' scores on the midtest (the middle column of Table 2), and the correlation in the CM Same condition was significantly higher than in the CM Different condition, $z = 2.09$, $p < .05$. Furthermore, the correlation is considered large. However, there was no correlation between tutors' rankings and students' midtest scores in the FTF Same (or the FTF Different) condition, and no difference between the FTF Same and FTF Different conditions, $z = 1.12$, $p > .10$. The correlation between tutors' rankings and students' actual competence was marginally significantly higher in the CM Same condition than in the FTF Same condition, $z = 1.50$, $p < .10$.

Table 2: Predicted versus actual student competence.

Condition	r (rankings)	r (# concepts)
CM Same	**+.69**[a]	**+.55**[b]
CM Different	**+.06**[c]	**-.16**[c]
FTF Same	+.27[c]	+.30[c]
FTF Different	-.14[c]	+.31[c]

[a] $p < .005$. [b] $p = .01$. [c] $p > .10$.

(2) *Tutors' assessments of students' relative physics competences: total number of concepts.* There was a significant positive correlation (again, large) between the total number of concepts tutors predicted their students would know on the midtest and the actual number of those concepts students demonstrated knowledge of in the CM Same condition, but not in the FTF Same, CM Different, or FTF Different conditions (the last column of Table 2). The correlation in the CM Same condition was significantly higher than in the CM Different condition, $z = 2.24$, $p = .01$, but there was no difference between FTF correlations, $z = 0.03$, $p > .10$. However, the correlation between the number of concepts tutors predicted their students would know and students' actual competence was not significantly higher in the CM Same than in the FTF Same condition, $z = 0.90$, $p > .10$.

(3) *Tutors' assessments of students' physics competence: absolute difference.* Tutors' assessments of their students' absolute competence levels were measured for each tutor/student pair as the absolute value of the difference between the number of concepts tutors predicted their students would know and the number of those concepts that students actually demonstrated knowing[6]. On this

accepted) suggests that tutors may not able to assess this type of student knowledge in a conceptual domain.

[4] Because the first three segments of the FTF conditions have not been transcribed but the all of final segments have been, the estimated number of words spoken by students in the first three segments was estimated by multiplying the average number of words per time spoken by students in the fourth segment by the average total time of the first three segments.

[5] The estimated proportion of words spoken to words typed was identical to the actual proportion of spoken to typed utterances reported in Condon and Cech (1996).

[6] Tutors significantly over-estimated the total number of concepts their students would answer correctly on the midtest in the CM Same condition, $t(19) = 4.70$, $p < .001$, and in the FTF Same condition, $t(19) = 3.53$, $p < .005$. However, the CM and FTF Same conditions did not differ, $t(38) = 0.66$, $p > .10$. That tutors over-estimate how much their students know is consistent with the result reported in Chi et al. (accepted).

measure, there were no differences between the CM Same and CM Different conditions, $t(38) = 1.15$, $p > .10$, between the FTF Same and FTF Different conditions, $t(38) = 0.49$, $p > .10$, or between the FTF Same and CM Same conditions, $t(38) = 0.05$, $p > .10$.

(4) *Tutors' assessments of students' knowledge of individual concepts.* To assess tutors' sensitivities to their students' knowledge of individual physics concepts, for all 20 tutor/student pairs in each condition, each prediction tutors made about whether their student would know a given concept on the physics midtest was compared with students' demonstrated knowledge of that concept on the physics midtest.

Table 3: Tutor sensitivities to individual knowledge.

	Loglinear results	
Condition	$?^2(2)$	$p(?^2)$
CM Same CM Different	**6.14**	**< .05**
FTF Same FTF Different	**9.40**	**< .01**

Tutors in the CM Same condition showed greater sensitivity to whether or not their students demonstrated knowledge of the concepts than tutors in the CM Different condition (Table 3). Similarly, tutors in the FTF Same condition showed greater sensitivity to their students' conceptual knowledge than tutors in the FTF Different condition. However, there was no difference between FTF Same and CM Same conditions, $?^2(3) = 0.41$, $p > .10$.

(5) *Tutors' assessments of students' motivation: correlational measure.* To measure the accuracy of tutors' assessments of their students' overall motivation, the correlation between each student's total score on the motivation questionnaire and the sum of the tutors' predicted student responses on the motivation questionnaire were computed for each condition (the middle column of Table 4). This measure may be considered a measure of tutors' sensitivity to students' relative overall motivation (including, for example, students' interest in physics, their confidence in their ability in physics). Neither correlation in the CM context was significant, and there was no difference between correlations in the CM context, $z = 0.24$, $p > .10$. The correlation in the FTF Same condition was marginally significant, whereas in the FTF Different condition the correlation was not significant; however, there was no significant difference in correlations, $z = 0.98$, $p > .10$. Nor was there a difference between the CM Same and FTF Same conditions, $z = 0.11$, $p > .10$.

(6) *Tutors' assessments of students' motivation: absolute measure.* As a more precise measure of the accuracy of tutors' assessment of their students' responses to statements on the motivation questionnaire, tutors' predictions of students' responses to individual statements on the motivation questionnaire were compared to students' actual responses. For each tutor/student pair, the absolute values of the differences between tutors' predictions of and students' responses to each statement on the motivation questionnaire were summed across all statements (Table 4, right column). Averages of these sums were compared across conditions. In the CM context, the mean sum in the CM Same condition was significantly lower than in the CM Different condition, $t(38) = 2.2$, $p < .05$. Similarly, in the FTF context, the mean sum in the FTF Same condition was significantly lower than in the FTF Different condition, $t(38) = 2.08$, $p < .05$. There was no difference across tutoring contexts, $t(38) = 0.00$, $p = 1$.

Table 4: Measures of motivation assessment accuracy.

Condition	r	mean (SD)
CM	$+.37^a$	**28.2 (10.08)**
CM Different	$+.30^a$	**33.9 (6.97)**
FTF	$+.40^b$	**28.2 (7.45)**
FTF Different	$+.08^a$	**35.3 (12.38)**

$^a p > .10$. $^b p < .10$.

(7) *Tutors' assessments of students' confidence and competitiveness: relative measure.* Though there were no differences between contexts in the accuracies of tutors' assessments of their students' overall motivation, perhaps there were differences between conditions in the specific types of motivational information tutors derive. Factor analysis on motivation questionnaire statements identified seven component factors, accounting for over 73% of the total variance. The first component factor, which comprised 33.75% of the total variance, loaded most highly on six statements related to students' confidence in physics (e.g., I know I will do well in the tutoring session and on the physics posttest today; I am generally good in physics). The second component factor, which comprised 11.34% of the total variance, loaded most highly on four statements related to competition with other students in physics (e.g., I try (or tried) to do better on physics exams than my friends; I would like being the only one who knew an answer to a physics question). The remaining component factors each comprised less than 10% of the total variance and will not be discussed here. The accuracies of tutors' assessments of their students' relative levels of confidence and competitiveness were assessed by first comparing the correlations between tutors' predictions of their students' total scores for all of the statements relating to the factor and students' total scores for those statements.

For the confidence dimension, though the correlation in the CM Same but not in the CM Different condition was significant, there was no significant difference between

these correlations, z = 0.62, p > .10 (Table 5, middle column). In the FTF conditions, neither correlation was significant, nor was there a significant difference in correlation between the FTF conditions, z = 0.50, p > .10. There was no difference between the CM Same and FTF Same conditions, z = 0.77, p > .10.

For the competitiveness dimension, neither correlation for the CM conditions was significant, nor was there a significant difference between correlations for the CM conditions, z = -0.58, p > .10. Similarly, for the FTF conditions, neither correlation was significant, nor was there a difference between conditions, z = -0.03, p > .10. There was no difference in correlation between the CM Same and FTF Same conditions, z = 0.77, p > .10.

Table 5: Correlations for motivation dimensions.

	Dimension	
Condition	confidence	competitiveness
CM Same	+.48[a]	.00
CM Different	+.30	+.20
FTF Same	+.25	+.25
FTF Different	+.08	+.26

[a] p < .05.

(8) *Tutors' assessments of students' confidence and competitiveness: absolute measure.* For each tutor/student pair, the absolute values of the differences between tutors' predictions of and students' responses to each statements loading on the dimension in question were summed across all statements loading on that dimension.

Table 6: Mean sum of absolute differences between tutors' predicted and students' actual response.

	Dimension	
Condition	confidence	competitiveness
CM Same	7.60 (2.95)	**3.05 (2.24)**
CM Different	7.8 (3.56)	**4.3 (2.43)**
FTF Same	*7.8 (3.19)*	2.25 (1.71)
FTF Different	*9.9 (4.80)*	2.8 (1.51)

Averages of these sums were compared across conditions as reported in Table 6. There was no difference between CM conditions in the accuracy of tutors' assessments of their students' confidence, $t(38) = 0.19$, p = .42. However, there was a marginally significant difference between FTF conditions, $t(38) = 1.63$, p = .06. There was no difference between the CM and FTF Same conditions, $t(38) = 0.21$, p = .84. For assessments of students' competitiveness, tutors in the CM Same condition were significantly more accurate than tutors in the CM Different condition, $t(38) = 1.69$, p < .05. However, there were no differences between the FTF conditions, $t(38) = 1.08$, p = .14, or between the CM and FTF Same conditions, $t(38) = 1.27$, p = .21.

Discussion

The results of this study showed that the accuracy of tutors' assessments of their students was dependent on the type of student information assessed (Table 7 summarizes results), and, to a lesser extent, the tutoring context.

Table 7: Summary of significance of results.

	Same versus Different:		CM vs. FTF:
Type of assessment	CM	FTF	Same
General competence			
Relative measures:			
(1) rankings	**Y**[a]	N	*Y*[b]
(2) total # concepts	**Y**	N	N
(3) Absolute measure	N	N	N
(4) Conceptual knowledge	**Y**	**Y**	N
Total Motivation:			
(5) correlational measure	N	N	N
(6) absolute measure	**Y**	**Y**	N
Motivational Dimensions:			
(7) correlational measure			
Confidence	N	N	N
Competitiveness	N	N	N
(8) absolute measure			
Confidence	N	*Y*	N
Competitiveness	**Y**	N	N

[a] Bolded font represents significant result.
[b] Italicized font represents marginally significant result.

In the CM context, there was evidence that tutors were able to assess their students' overall relative physics competence, measured both as tutors' rankings of their students' general competence on a 7-point Likert scale (from 1 "well below average" to 7 "well above average") and as the sum of the number of concepts tutors predicted their students would answer correctly on the midtest. Each tutor predicted the general competence of one student only and there was a significant correlation; one possible explanation for how tutors were able to gauge the relative competence of their student even though they did not have for comparison other tutored students is that tutors have a common conception of the general competence of an "average" student, which CM tutors were able to compare their students against. There was no evidence that tutors were able to accurately assess their students' overall relative competence by either measure in the FTF context.

Though tutors may have developed more accurate assessments of their students' relative general physics competences in the CM context, tutors in both contexts seemed to have developed about equally accurate assessments of their students' conceptual knowledge (i.e., which concepts their students did and did not know).

There was no evidence that tutors in either context developed accurate assessments of the various dimensions of motivation assessed by the questionnaire (the correlational analysis). This was also true for two dimensions of student motivation: confidence and competitiveness.

Overall, the only result for which there was a context effect was for tutors' assessments of their students' relative general competence, where tutors in the CM condition were marginally more accurate. Perhaps surprisingly, there was no strong evidence that tutors developed more accurate assessments of their students' overall motivation, confidence, or competitiveness in a FTF context, where more sources of information about students' affective states are available. On the whole, these results are encouraging to designers of intelligent tutoring systems, suggesting the possibility that at least equally accurate assessments of students may be possible through text-only communication[7].

Acknowledgements

This research was supported by the Office of Naval Research, Cognitive and Neural Sciences Division MURI Grant N00014-00-1-0600 and NSF Grant 9720359 to CIRCLE, a center for research on intelligent tutoring.

We also thank Susan Fussell, Jonathan Schooler, and Chris Schunn for their many insightful comments, especially during the earlier stages of the project. Thanks to Marguerite Roy for her help with statistics (though any errors are the sole responsibility of the first author), and thanks to Chas Murray, H. Chad Lane, Noboru Matsuda, and Min Chi for their comments on this paper.

References

Bloom, B. S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher*, 4-16.

Chi, M. T. H. (1996). Constructing self-explanations and scaffolded explanations in tutoring. *Applied Cognitive Psychology, 10 (Special Issue)*, 33-49.

Chi, M. T. H., Siler, S. A. & Jeong, H. (Accepted). Do tutors accurately monitor students' understanding? *Cognition and Instruction*.

Clark, B., Bratt, E. O., Lemon, O., Peters, S., Pon-Barry, H. Thomsen-Gray, Z. & Treeratpituk (2002). *A General Purpose Architecture for Intelligent Tutoring Systems*. In the proceedings of the International CLASS Workshop on Natural, Intelligent, and Effective Interaction in Multimodal Dialogue Systems.

Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. In L. B. Resnick, J. M. Levine, & S. D. Teasley (Eds.), *Perspectives on Socially Shared Cognition* (pp. 127-149). Washington, DC: American Psychological Association.

Cohen, P. A., Kulik, J. A., & Kulik, C. C. (1982). Educational outcomes of tutoring: A meta-analysis of findings. *American Educational Research Journal, 19*(2), 237-248.

Condon, S. L. & Cech, C. G. (1996). Functional comparison of face-to-face and computer-mediated decision making interactions. In S. C. Herring (Ed.), *Computer-mediated communication* (pp. 65-80). Philadelphia, PA: John Benjamins.

Fox, B. A. (1993). *The Human Tutorial Dialogue Project*. Hillsdale, NJ: Erlbaum.

Hausmann, R. G. M., & Chi, M. T. H. (2002). Can a computer interface support self-explaining? *Cognitive Technology, 7*(1), 4-15.

Lebie, L., Rhoades, J. A., & McGrath, J. E. (1996). Interaction process in computer-mediated and face-to-face groups. *Computer Supported Cooperative Work, 4*, 127-152.

Litman, D. J. (2002). *Adding Spoken Dialog to a Text-Based Tutorial Dialog System.* In the proceedings of the ITS2002 Workshop on Empirical Methods for Tutorial Dialogue Systems.

Ruberg, L. F., Moore, D. M., & Taylor, C. D. (1996). Student participation, interaction, and regulation in a computer-mediated communication environment: A qualitative study. *Journal of Educational Computing Research, 14*(3), 243-268.

Wigfield, A., & Guthrie, J. T. (1997). Relations of children's motivation for reading to the amount and breadth of their reading. *Journal of Educational Psychology, 89*, 420-433.

[7] A question that is still open and will be investigated is the amount of information tutors received from the students' actions in the drawing window.