Patterns of Word Usage in Expert Tutoring Sessions: Verbosity versus Quality

Sidney D'Mello

Institute for Intelligent Systems, 202 Psychology Building, University of Memphis, Memphis TN, 38152, USA
sdmello@memphis.edu

Abstract

It is widely acknowledged that one-on-one human tutoring is one of the most effective ways to provide learning, however, the source of its effectiveness is still unclear. Tutor-centered, student-centered, and interaction hypotheses have been proposed as possible explanations of the effectiveness of human tutoring. Most research has addressed this question by analyzing tutorial sessions at the dialogue move or speech act level. The present paper adopts a different approach by focusing on word usage patterns in 50 naturalistic tutorial sessions between human students and expert tutors. Specifically, each unique word in the session was designated as a student initiative word, a tutor initiative word, or a shared-initiative word. Comparisons of the frequencies as well as the weights of the words assigned to each of these categories indicated that the student and tutor share initiative even though the tutor's are considerably more verbose. The implications of the results for the development of an ITS that aspires to model expert tutors are discussed.

Introduction

There is no one-size-fits-all approach to learning and instruction. It only takes a few probing questions with Socratic dialogues to effectively teach gifted students, while substantial direct instruction and detailed explanations are needed for less knowledgeable learners (D'Mello, Hays et al., 2010). Intrinsically motivated learners derive pleasure from the task itself (e.g., enjoyment from problem solving), while learners with extrinsic motivation rely on external rewards (e.g., receiving a good grade) (Elliot & McGregor, 2001). Adventuresome learners want to be challenged with difficult tasks, take risks of failure, and manage negative emotions when they occur, while cautious learners tackle easier tasks, take fewer risks, and minimize failure and its resulting negative emotions (Clifford, 1988).

Unfortunately, most classrooms do not afford pedagogical interventions that are tailored at the individual student level, so it comes as no surprise that many students fail and fall behind. These students may turn to one-on-one human tutoring when they are having difficulty in courses. Investing time and effort in one-on-one tutoring does have a big payoff, as evident from the substantial empirical evidence showing that human tutoring is extremely effective when compared to typical classroom environments (Bloom, 1984; Cohen, Kulik, & Kulik, 1982; Corbett, 2001).

The effectiveness of one-on-one tutoring in human and computer tutors raises the question of *what* makes tutoring so powerful? This is a pertinent question because understanding the tactics and dialogue moves of human tutors has direct application for Intelligent Tutoring Systems (ITSs), especially those that aspire to model human tutors (D'Mello, Hays et al., 2010; VanLehn, 2006). Although ITSs are quite effective in promoting learning gains, and even outperform some human tutors (Corbett, 2001; Dodds & Fletcher, 2004; VanLehn et al., 2007), there is still room for improvement. This presents the challenge of better understanding human tutoring with an eye for implementing important insights into next generation ITSs, which is the goal of the present paper.

What Makes Human Tutoring Effective?

Chi and colleagues formulate three different hypotheses, namely the *tutor-centered, student-centered, and interaction hypotheses*, as possible explanations of the effectiveness of human tutoring (Chi, Roy, & Hausmann, 2008). The tutor-centered hypothesis contends that it is the pedagogical and motivational strategies of the tutor that underlie the effectiveness of one-on-one tutoring. Alternatively, the student-centered hypothesis predicts that tutoring is effective because it gives students more opportunities to actively construct knowledge, rather than anything the tutor does in particular. Lastly, the interaction hypothesis is essentially the blending of the tutor and student-centered hypotheses, with a focus on the coordinated effort of both tutor and student.

The tutor-centered hypothesis has been of primary focus over the past several decades of research, yielding some important insights about the pedagogical strategies employed by tutors. For example, research has shown how tutors adapt to students needs by (a) modeling and monitoring student knowledge (Chi, Siler, Jeong, Yamauchi, & Hausmann, 2001), (b) employing tutoring tactics and strategies that contribute to learning gains (Lajoie, Faremo, & Wiseman, 2001), (c) planning at local and global levels of discourse (Littman, Pinto, & Soloway, 1990), and (d) providing emotional support for students in social, affective, and motivational ways (del Soldato & du Boulay, 1995; Lepper, Woolverton, Mumme, & Gurtner, 1993).

The student-centered hypothesis contends that students are active participants in the construction of their own knowledge, rather than being mere information receptacles. This hypothesis has found substantial support in the tutoring literature. For example, considerable evidence suggests that constructive student moves such as self explanations (Chi, Deleeuw, Chiu, & Lavancher, 1994) and question asking (Graesser & Olde, 2003) are beneficial to learning.

In contrast, the interaction hypothesis predicts that the effectiveness of tutoring draws from both tutor and student behavior and their *coordination* with each other. The interaction hypothesis has substantial overlap with the collaborative learning literature where it has been shown that group learning outperforms individual learning (Wiley & Bailey, 2006). Simply put, both stress the importance of the interaction between participants.

Research Goals

Investigations into the nature of one-on-one tutoring will inevitably encounter issues pertaining to "grain size" or the level of analysis required to answer certain theoretical questions. The analysis of tutorial dialogue usually takes place at a very fine-grained level, or the speech act or dialogue move level (e.g., hints, prompts, positive feedback) (D'Mello, Lehman, & Person, 2010; Graesser, Person, & Magliano, 1995). More recently, there has been an emphasis on analyzing larger chunks of dialogue moves (called dialogue modes) that span a few minutes and provide the overarching context or teaching phase (e.g., a lecture, a problem scaffolding phase) during which learning occurs (Boyer et al., 2009; Cade, Copeland, Person, & D'Mello, 2008). Yet another possibility is to bridge the gap between the move and mode level by analyzing clusters of dialogue moves within a dialogue mode (D'Mello, Olney, & Person, in press).

The present paper adopts a different approach. 50 naturalistic tutorial sessions between human students and expert tutors were analyzed by focusing on word usage patterns over the course of the session. Specifically, each unique word was designated as a student initiative word, a tutor initiative word, or a shared-initiative word. This classification allows us to investigate who takes initiative in the tutoring session, thereby affording a comparison of the tutor-centered, student-centered, and interaction hypotheses. For example, shared-initiative is low if both conversational partners use a significant amount of unique words in a window of turns (i.e., there is negligible overlap in the words uttered by both student and tutor). The advantage of the current approach is that it might support the automated identification of initiative taking without the need to perform tedious annotations for speech acts, dialogue moves, and modes.

The present focus is on expert human tutoring sessions because it is widely acknowledge that expert tutors are very effective at promoting learning gains and motivating students (Bloom, 1984; Lepper & Woolverton, 2002). Any insights obtained from this analysis of expert human tutoring can be used to guide the development of ITSs that model expert tutors.

Expert Tutoring Corpus

Data Collection

The corpus consisted of 50 tutoring sessions between 39 students and 10 expert tutors on the subjects of algebra, geometry, physics, chemistry, and biology. The students were all having difficulty in a science or math course and were either recommended for tutoring by school personnel or voluntarily sought professional tutoring help.

The expert tutors were recommended by academic support personnel from public and private schools in a large urban school district. All of the tutors had long-standing relationships with the academic support offices that recommended them to parents and students. The criteria for being an expert tutor were: (a) have a minimum of five years of one-to-one tutoring experience, (b) have a secondary teaching license, (c) have a degree in the subject that they tutor, (d) have an outstanding reputation as a private tutor, and (e) have an effective track record (i.e., students who work with these tutors show marked improvement in the subject areas for which they receive tutoring).

Fifty one-hour tutoring sessions were videotaped and transcribed. There were 31 sessions on math topics (algebra and geometry) and 19 sessions on science topics (physics, chemistry, and biology). A total of 16,728 student-tutor dialogue turns (or simply turns) were extracted in the 50 hours of tutoring. The number of turns per session ranged from 113 to 752 with each session containing an average of 334 turns (*SD* = 136). The number of unique words per session ranged from 378 to 1015 with an average of 754 words (*SD* = 163).

Data Annotation and Scoring

The corpus was preprocessed in order to eliminate meta tags and punctuation. Next, unique words in each session were identified and assigned to one of five categories: (1) student unique, (2) tutor unique, (3) student lead, (4) tutor lead, and (5) student-tutor align. The first two categories represent specific vocabulary words uttered by either the student (i.e., student unique) or the tutor (i.e., tutor unique), but never by both conversation participants. Words in groups 3-5 (i.e., student lead, tutor lead, and student-tutor align) were uttered by both participants, thereby representing a shared vocabulary between student and tutors. The critical discriminating feature for these words was the source and timing of the first utterance (i.e., who first introduced the word to the session). Specifically, a word uttered by both student and tutor in the same or adjacent student-tutor dialogue turns was assigned to the student-tutor align category, irrespective of who uttered the word first. Words in the student lead category were used by the student before the tutor, while it was the tutor who first used the common word in the tutor lead category. Categories 1 and 3 represent student-initiative words, 2 and 4 are tutor-initiative words, while words in category 5 are shared- or mixed-initiative words.

All analyses were conducted at the session level. Two dependent measures (i.e., proportional scores and weighted scores) were computed for each category, yielding 10 measures in all. The first five measures consisted of the proportional assignment of words to each of the five categories (the five proportional scores for each session add up to 1). While these proportional measures represent the distribution of words in each category, the second set of measures was sensitive to the quality of words in each category. Specifically, each word in the corpus was weighted on a scale from 0.0 to 1.0, relative to its inverse frequency in the English language using the CELEX corpus (Baayen, Piepenbrock, & Gulikers, 1995). As a consequence, higher frequency words such as closed-class function words (e.g., and, but) have comparatively lower weights than lower frequency words (e.g., mitosis, Newtonian) that have higher weights. These lower frequency words are generally domain-related content words while the high frequency words are function and domain-independent content words (e.g., should, calculator). For example, “the,” which is the most common word in the English language, has a weight of .05, while extremely rare terms like “mitosis,” and “muzzle-velocity” have weights of 1.0.

Weighted scores were computed by averaging the weights of the words in each category. In this fashion, both the frequency (proportional scores) and importance (weighted scores) of each category can be compared.

Results and Discussion

Patterns of Word Usage across Entire Session

Table 1 provides descriptive statistics on proportional scores and weighted scores for each category. It also lists a correlation between proportional scores and weighted scores. With the exception of the student-tutor align category, proportional and weighted scores were not significantly correlated, so these measures captured unique aspects of how words are being used in the tutoring sessions. Importantly, the medium-sized (Cohen, 1992) positive correlation between proportional and weighted scores for the student-tutor align category suggest that as the degree of alignment increases, so do the weights of the aligned words.

Two 5 × 2 (category × topic) mixed ANCOVA were conducted to assess if there were statistical differences in proportional and weighted scores of the five categories. Category was a within-subjects factor with 5 levels for proportional scores associated with the five word categories. Topic was between-subjects factor with two levels for math ($N = 31$) and science ($N = 19$). Topic was included in the model to assess whether the subject of the tutorial session had an impact on how words were being used. The number of turns and words in each session were added as covariates in order to control for any spurious effects that might be attributed to these variables.

Proportional Scores. The results indicated that there was a significant main effect for category, $F(4, 184) = 10.1$, $Mse = .006$, $p < .001$, partial eta-square = .179. The main effect for the topic factor was not tested because it was constrained since proportional scores within each topic sum up to 1. The category × topic interaction was not significant ($p = .609$), so the tutorial topic had no noticeable impact on the proportional scores.

Bonferroni posthoc tests on the category main effect, revealed the following pattern in the data at the $p < .05$ level: *tutor unique > tutor lead > student unique > student lead > student-tutor align*. Hence, if one simply counts the words in each category, it appears that it is the tutor who takes the initiative by using unique words and leading the students. In fact, more than half of the words (54.5%) uttered in the tutorial session were part of the tutor’s unique vocabulary. An additional 18.4% of the words were first introduced by the tutor and subsequently used by the student. Taken together, the tutor takes initiative for 72.9% of the words, the student takes initiative for 22.3% of the words, and, surprisingly, a mere 4.5% of the words were shared-initiative words.

Table 1. Patterns of word usage by students and tutors

	Proportions		**Weights**		
Measure	*M*	*SD*	*M*	*SD*	*r*
S Unique	.123	.075	.141	.021	.240
T Unique	.549	.132	.134	.009	-.082
S Lead	.100	.049	.103	.015	.178
T Lead	.184	.037	.102	.008	-.022
ST Align	.045	.016	.133	.030	.372**

Note. S = student, T = tutor, ST = student-tutor, ** $p < .05$

Weighted Scores. The ANCOVA on weighted scores indicated that there was a significant main effect for category, $F(4, 184) = 4.24$, $Mse = .0003$, $p = .003$, partial eta-square = .084. Bonferroni posthoc tests revealed the following pattern at the $p < .05$ level: *(tutor unique = student-tutor align) > (student lead = tutor lead).* So, the weighted scores associated with the tutor unique words and the student-tutor align words were on par and significantly greater than both the student and tutor lead words, which were equivalent to each other. The student unique words fit this general pattern, except that the mean score associated with these words was significantly greater than the mean score for the tutor unique words. These results signal an important difference from the patterns obtained with the proportional scores because it is no longer the tutor who takes most of the initiative.

In addition to the significant main effect for category, the category × topic interaction was also statistically significant, F(4, 184) = 4.42, Mse = .0003, p = .002, partial eta-square = .088. Posthoc tests indicated that weighted scores for student unique words were higher for science (M = .152, SD = .028) compared to math (M = .135, SD = .012); the difference was consistent with a large effect (d = .79). Conversely, tutor unique weighted scores were greater for math (M = .136, SD = .009) than science (M = .132, SD = .009), d = .44. There were no topic differences for lead and alignment words.

Verbosity vs. Quality

The results so far paint a mixed picture of initiative-taking in expert tutoring. The proportional scores indicate that it is the tutor who does most of the talking; however, the weighted scores suggest that both the student and tutor share the initiative in the session. In order to reconcile between these diverging sets of results, a follow-up analysis that compared proportional and weighted scores for each category was performed. Since the two scores are measures of intrinsically different constructs (verbosity vs. quality), all scores were standardized prior to the analyses.

The analyses proceeded with a 5 × 2 × 2 (category × measure × topic) mixed ANCOVA. Measure was a within-subjects factor with two levels for proportional and weighted scores. As before, the number of turns and words in each session were included as covariates. Of greatest interest is the category × measure interaction, which was significant, *F*(4, 176) = 5.36, *Mse* = .332, *p* = .003, partial eta-square = .109.

Three important conclusions can be gleaned from Bonferroni posthoc tests on the category × measure interaction presented in Figure 1. First, the difference between proportional and weighted scores for both student and tutor unique words was significant. Effect sizes were 1.76 and 2.43 sigma for student and tutor unique words, respectively. These large effects indicate that although students use fewer unique words when compared to tutors, student unique words have higher weights than tutor unique words.

Second, proportional scores were significantly greater than weighted scores for both student (*d* = .58) and tutor (*d* = 2.99) lead words. The medium effect size for student lead words and the large effect size for tutor lead words support the general pattern that the verbosity-quality difference is more pronounced for the tutors then for the students.

The third important finding pertains to the student-tutor align category. There was a large difference between proportional scores and weights associated with these words (*d* = 1.38). Hence, although students and tutors rarely use the same words in the same turn or across adjacent turns, the words that they align on are presumably rare domain-related content words.

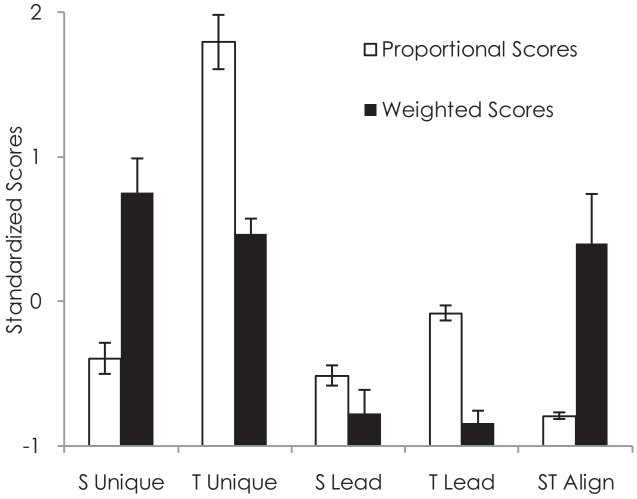

Figure 1. Proportional vs. weighted scores

Predicting Weighted Student-Tutor Alignment

The possibility of predicting weighted student-tutor alignment scores from the weighted unique and lead scores was considered. It was not possible to predict proportional alignment scores due to severe multicollinearity problems among proportional unique and lead scores. The analysis consisted of regressing student-tutor weighted scores on student and tutor unique and lead weighted scores. Separate models were constructed with either student or tutor scores as independent variables, therefore affording us with the ability to assess whether it is the student or tutor words that best predict alignment weights.

There was the concern that variability in topics, number of turns, and number of words would unduly influence the models. This concern was addressed by conducting two-step multiple regression models. Step 1 predictors consisted of a dummy coded variable for topic (math = 1; science = 0), the number of turns, and the number of words in each session. The Step 2 predictors were the weighted scores of the unique and lead categories for the student or tutor. The Step 2 models predict residual variance above and beyond the Step 1 models. Hence, a significant Step 2 model would be indicative of predicting alignment scores after controlling for topic, number of turns, and number of words.

Outliers (values greater than 2*SD* from the mean) were identified and removed prior to constructing the models. A tolerance analysis indicated that there were no critical multicollinearity problems since all predictors had tolerance values greater than 0.5; tolerances above 0.4 are recommended (Allison, 1999).

The analysis yielded a significant Step 2 model when weighted scores for unique and lead student words (but not the tutor words) were the predictors. The overall student model was significant, $F(5, 38) = 3.54$, $p = .01$ and had an adjusted R^2 of .228; this is consistent with a medium to large effect (Cohen, 1992). The Step 2 model explained an additional 9.2% of the variance over the Step 1 model. Hence, it is the words that students use that are most predictive of student-tutor alignment.

The parameters of the multiple regression model are presented in Table 2. It appears that the weighted scores associated with unique words used by the student significantly predict student-tutor alignment weights.

Table 2. Parameters of multiple regression model

Parameter	*B*	*β*	*t*	*p*
Intercept	.028		.627	.534
Topic	-.015	-.325	-2.07	.045
No. Turns	.000	.059	.407	.686
No. Words	.000	-.087	-.568	.574
S Unique Weighted Score	**.596**	**.342**	**2.31**	**.027**
S Lead Weighted Score	.303	.159	1.10	.279

General Discussion

The word-level analysis of a large corpus of expert tutoring sessions yielded some important conclusions about tutor-student initiative-taking. Tutors were substantially more verbose as approximately three-fourths of the distinct words in the corpus were tutor-initiative words. The strength of this effect might make it tempting to accept the tutor-initiative hypothesis, which posits that it is primarily the tutor's actions that underlie the effectiveness of one-on-one tutoring.

The weighted scores, however, preclude us from accepting this hypothesis too cavalierly. The analysis that focused on word weights, instead of mere frequencies, indicated that the student and tutor share initiative even though the tutors are considerably more verbose. Simply put, tutors are more verbose while students are more selective.

An analysis of student speech acts does shed some light on the low student verbosity. When a 16 category coding scheme was also applied to classify student dialogue moves (D'Mello et al., in press), the results indicated that students primarily spoke in response to a tutor question or to provide back-channel feedback. In fact, 63% of student moves in the corpus consist of conversational acknowledgements or responses to tutor questions. Hence, one explanation for the increased tutor verbosity is that they need to be doing most of the talking in order to keep the conversation flowing. This sketch is intuitively plausible because according to Graesser et al. (1995) it is the natural language dialogue patterns, as opposed to sophisticated pedagogy, that best explains the effectiveness of novice human tutoring. Might the same hold true for expert human tutoring?

At first blush, this claim is incompatible with the current view that expert tutors rely on sophisticated pedagogical and motivational strategies that are currently not on the radar of novice tutors and ITSs (Graesser et al., 1995; Lepper, Drake, & O'Donnell-Johnson, 1997; Lepper & Woolverton, 2002). However, it might be a combination of both sophisticated pedagogy and carefully nuanced dialogue management that underlie the effectiveness of expert tutoring.

The research team is currently in the process of developing a tutoring system (Guru) for high school biology based on the tactics, actions, and dialogue of expert human tutors. The pedagogical and motivational strategies of Guru are informed by a detailed computational model of expert human tutoring. In addition to refining and reconceptualizing the current understanding of expert human tutoring, this analysis of the dynamics of initiative taking will directly guide the behavior of Guru. Whether this expert tutor based ITS yields substantial benefits over current ITSs awaits further technological development and empirical testing.

Acknowledgments

The author gratefully acknowledges Natalie Person for providing the expert tutoring corpus. This research was supported by the National Science Foundation (ITR 0325428, HCC 0834847) and Institute of Education Sciences (R305A080594). Any opinions, findings and conclusions, or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the funding agencies.

References

Allison, P. D. (1999). *Multiple regression*. Thousand Oaks, CA: Pine Forge Press.

Baayen, R., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX lexical database (CD-ROM)*. Philadelphia: University of Pennsylvania.

Bloom, B. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher, 13*(6), 4-16.

Boyer, K., Young, E., Wallis, M., Phillips, R., Vouk, M., & Lester, J. (2009). Discovering Tutorial Dialogue Strategies with Hidden Markov Models. In V. Dimitrova, R. Mizoguchi, B. Du Boulay & A. Graesser (Eds.), *Proceedings of the 14th*

International Conference on Artificial Intelligence in Education (pp. 141 - 148). Amsterdam: IOS Press.

Cade, W., Copeland, J., Person, N., & D'Mello, S. (2008). Dialogue modes in expert tutoring. In B. Woolf, E. Aimeur, R. Nkambou & S. Lajoie (Eds.), *Proceedings of the 9th international conference on Intelligent Tutoring Systems* (pp. 470-479). Berlin, Heidelberg: Springer-Verlag.

Chi, M., Deleeuw, N., Chiu, M., & Lavancher, C. (1994). Eliciting Self-Explanations Improves Understanding. *Cognitive Science, 18*(3), 439-477.

Chi, M., Roy, M., & Hausmann, R. (2008). Observing tutorial dialogues collaboratively: Insights about human tutoring effectiveness from vicarious learning. *Cognitive Science, 32*(2), 301-341.

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

Clifford, M. (1988). Failure tolerance and academic risk-taking in ten- to twelve-year-old students. *British Journal of Educational Psychology, 58*(15-27).

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155-159.

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

Corbett, A. (2001). Cognitive computer tutors: Solving the two-sigma problem. In M. Bauer, P. Gmytrasiewicz & J. Vassileva (Eds.), *Proceedings of 8th International Conference on User Modeling* (pp. 137-147). Berlin / Heidelberg: Springer.

D'Mello, S., Hays, P., Williams, C., Cade, W., Brown, J., & Olney, A. (2010). Collaborative Lecturing by Human and Computer Tutors In J. Kay & V. Aleven (Eds.), *Proceedings of 10th International Conference on Intelligent Tutoring Systems* (pp. 609-618). Berlin / Heidelberg: Springer.

D'Mello, S., Olney, A., & Person, N. (in press). Mining collaborative patterns in tutorial dialogues. *Journal of Educational Data Mining, 2*(1), 1-37.

D'Mello, S. K., Lehman, B., & Person, N. (2010). Expert Tutors Feedback is Immediate, Direct, and Discriminating. In C. Murray & H. Guesgen (Eds.), *Proceedings of the 23rd Florida Artificial Intelligence Research Society Conference* (pp. 595-560). Menlo Park, California: AAAI Press.

del Soldato, T., & du Boulay, B. (1995). Implementation of motivational tactics in tutoring systems. *International Journal of Intelligence in Education, 6*, 337-378.

Dodds, P., & Fletcher, J. (2004). Opportunities for new "smart" learning environments enabled by next-generation web capabilities. *Journal of Educational Multimedia and Hypermedia, 13*(4), 391-404.

Elliot, A., & McGregor, H. (2001). A 2 x 2 achievement goal framework. *Journal of Personality and Social Psychology, 80*(3), 501-519.

Graesser, A., & Olde, B. (2003). How does one know whether a person understands a device? The quality of the questions the person asks when the device breaks down. *Journal of Educational Psychology, 95*(3), 524-536.

Graesser, A., Person, N., & Magliano, J. (1995). Collaborative dialogue patterns in naturalistic one-to-one tutoring. *Applied Cognitive Psychology, 9*(6), 495-522.

Lajoie, S., Faremo, S., & Wiseman, J. (2001). Tutoring strategies for effective instruction in internal medicine. *International Journal of Artificial Intelligence and Education, 12*, 293-309.

Lepper, M., Drake, M., & O'Donnell-Johnson, T. (1997). Scaffolding techniques of expert human tutors. In K. Hogan & M. Pressley (Eds.), *Scaffolding student learning: Instructional approaches and issues* (pp. 108-144). New York: Brookline Books.

Lepper, M., & Woolverton, M. (2002). The wisdom of practice: Lessons learned from the study of highly effective tutors. In J. Aronson (Ed.), *Improving academic achievement: Impact of psychological factors on education* (pp. 135-158). Orlando, FL: Academic Press.

Lepper, M., Woolverton, M., Mumme, D., & Gurtner, J. (1993). Motivational techniques of expert human tutors: Lessons for the design of computer based tutors. In S. Lajoie & S. Derry (Eds.), *Computers as cognitive tools*. Hillsdale, NJ: Erlbaum.

Littman, D., Pinto, J., & Soloway, E. (1990). The knowledge required for tutorial planning: An empirical analysis. *Interactive Learning Environments, 1*, 124-151.

VanLehn, K. (2006). The behavior of tutoring systems. *International Journal of Artificial Intelligence in Education., 16*(3), 227-265.

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

Wiley, J., & Bailey, J. (2006). Effects of collaboration and argumentation on learning from web pages. In A. M. O'Donnel, C. E. Hmelo-Silver & G. Erkens (Eds.), *Collaborative learning, reasoning, and technology* (pp. 297-321). Mahwah, New Jersey:: Lawrence Earlbaum.