Working With Pedagogical Agents: Understanding the “Back End” of an Intelligent Tutoring System

Christopher R. Wolfe
Colin L. Widmer
Audrey M. Weil
Elizabeth M. Cedillos-Whynott
Miami University

Students in an undergraduate psychology course on Learning and Cognition used SKO (formerly AutoTutor Lite), an Intelligent Tutoring System, to create interactive lessons in which a pedagogic agent (animated avatar) engages users in a tutorial dialogue. After briefly describing the technology and underlying psychological theory, data from an anonymous student survey and confidential focus group are presented. Results indicate that students successfully met established criteria for the assignment. Students rated the assignment highly and reported high levels of understanding resulting from the experience. Results are discussed in the context of learning experiences to help students gain an intuitive feel for complex phenomena.

“Siri, where can I get a beer around here?” I (the first author) started class by addressing my iPhone rather than my students. “I found 15 bars fairly close to you,” replied Siri, a talking smart phone application. There are several noteworthy aspects of this verbal exchange between me and this computer-based natural language processing system, as soon became apparent in our class discussion. Obviously, Siri had to recognize my voice and “understand” my speech. Moreover, Siri had to “know” that beer is found in bars—the kind of inference that we typically take for granted when talking with one another. Finally, Siri had to know where I was (my physical location) in relation to a list of local taverns it provided, including the name, address, distance, and customer rating for each.
In recent years there has been an explosion in the availability and sophistication of such “intelligent” discourse technologies. Although talking technologies have been around for quite some time, machines capable of understanding what we say, making appropriate inferences, and responding in appropriately useful ways have only recently moved from the lab to everyday use.

A current challenge put forth by President Obama in his April 2009 speech to the National Academy of Science is to create “learning software as effective as a personal tutor” (Anissmov, 2009). One promising tack is to develop learning systems that feature some of the same processes that lead to effective learning in one-on-one human tutoring, which is the goal of most Intelligent Tutoring Systems (ITS). These ITS engage in human-computer interactions that are meant to simulate the experience of a student talking with a human tutor (Graesser et al., 2004). I wanted my cognitive psychology students to develop a good understanding of how ITS work, which is the primary goal of the exercise described below.

**Learning & Cognition**

The course Psychology 372: Learning & Cognition at Miami University in Oxford, Ohio was the context for this learning experience. The course focuses on human learning with an emphasis on conceptual learning rather than operant or classical conditioning. There were 22 upper-division undergraduate students about equally split between psychology majors and non-majors taking the course to fulfill requirements. The course considered a range of psychological research and theory and a variety of settings including university courses in chemistry and history, informal education, and learning from games.

Learning & Cognition is a course I have taught four times, and in recent years I have included a brief module on tutoring and ITS. One-on-one tutoring is arguably the “gold standard” for human learning (Chi, Siler, Jeong, Yamauchi, & Hausmann, 2005). Trained human tutors reportedly produce gains as high as two standard deviations over standard classroom practice, sometimes labeled the “2 sigma effect,” though a recent literature review suggests effect sizes of about 0.79 are typical (VanLehn, 2011). Tutors engage students’ attention, ask questions, and give them immediate feedback on their progress (Graesser & McNamara, 2010). Advanced learning technologies such as ITS attempt to replicate the strong learning gains attained by well-trained human tutors.

Good human tutors typically encourage students to elaborate on their answers to knowledge questions (Chi et al., 2005). Research suggests that
actively generating and elaborating explanations of material is more beneficial to learning than passively spending time with the material by reading or listening to lectures (Graesser, McNamara, & VanLehn, 2005). Thus, encouraging elaboration and self-explanation is a key goal of many ITS.

ITS are psychologically interesting because they necessitate putting cognitive learning theory into practice. Moreover, those creating ITS must be explicit about all of their assumptions about how people learn and the nature of tutorial interactions. Thus they provide an informative lens through which to view the cognitive processes of deep conceptual learning. In the past, I have had students read and discuss peer-reviewed papers about ITS, and in this class we read and discussed papers by Arthur Graesser, 2011 winner of the American Psychological Association Award for Distinguished Contributions of Applications of Psychology to Education and Training, and colleagues on tutorial dialogues, learning with discourse technologies, and the emotions of learning (Baker, D’Mello, Mercedes, Rodrigo, & Graesser, 2010; D’Mello, Dowell, & Graesser, 2011; Graesser, 2011; Graesser et al., 2004;). Thanks to a grant from the National Cancer Association (Wolfe et al., Manuscript submitted for publication), I also had the opportunity to permit Learning & Cognition students to work directly with computer-based pedagogical agents themselves to help them understand the inner workings (that is, the “back end”) of an ITS.

**Intelligent Tutoring Systems**

One of the most successful ITS is AutoTutor (Graesser et al., 2004). AutoTutor is built around the psychological insight that when people actively generate explanations and justify their answers, learning is more effective and deeper than when they are simply given information (Arnott, Hastings, & Allbritton, 2008). The explanations are pedagogically deep because users must learn to express causal and functional relationships rather than mechanically applying procedures (Graesser et al., 2005).

AutoTutor has been applied to teaching students in many knowledge domains, including behavioral research methods (Arnott et al., 2008), computer science, (Craig, Sullins, Witherspoon, & Gholson, 2006; Graesser et al., 2004), and physics, (Jackson, Ventura, Chewle, Graesser, & the Tutoring Research Group, 2008 [2004 in references.]; VanLehn et al., 2006). In AutoTutor, an animated talking avatar or embodied conversational agent (Louwerse, Graesser, McNamara, & Lu, 2008) communicates with voice inflection, conversational phrasing, facial expressions, and simulated facial movements. Didactic information is communicated with graphical displays, including video with sound.
Sharable Knowledge Object (SKO, formerly AutoTutor Lite; Hu, Han, & Cai, 2008; Hu & Martindale, 2008; Wolfe, Fisher, Reyna, & Hu, 2012; Wolfe et al., under review) is a web-based version of AutoTutor developed by Dr. Xiangen Hu. SKO allows developers to create effective tutorial dialogues without a team of highly experienced computer scientists, as is necessary to develop dialogues in other ITS (Wolfe et al., 2013). It is one of the first web-based ITS, meaning that students can interact with it from anywhere using an ordinary browser. Like AutoTutor, SKO has a talking animated agent that interacts with users based on expectations using hints and elaboration. SKO presents learners with images, sounds, text, and videos. To respond appropriately to learner input, SKO compares text entered by students to a set of expectation texts using Latent Semantic Analysis (LSA; Hu, Cai, Wiemer-Hastings, Graeser, & McNamara, 2007). Like human tutors and other ITS, SKO elicits verbal responses from learners and encourages them to further elaborate their understanding. SKO can, thus, be used to encourage self-explanation (Roscoe & Chi, 2008; VanLehn et al., 2007).

Through a natural language dialogue with the learner, SKO guides the learner toward a set of target expectations. With SKO, tutorials are built from units called SKOs (Sharable Knowledge Objects; Hu, 2013). Each SKO presents materials to the learner didactically and then solicits a verbal response from the learner.

LSA is a computational technique that mathematically measures the semantic similarity of sets of texts (Hu et al., 2007). It accomplishes this by creating a semantic space from a large corpus of text. The semantic space is a representation of the semantic relations of words based on their co-occurrences in the corpus (Hu et al., 2007). In the context of an Intelligent Tutoring System, LSA is used to compare sentences entered by students to a specially prepared text that embodies good answers. The tutor can measure how close the user’s response is to the expectations text, give appropriate feedback, and encourage elaboration and other verbal responses based on this comparison (Kopp, Britt, Millis, & Graesser, 2012).

A detailed understanding of the mechanics of LSA is not required to use LSA in SKO. What is important is to have a basic understanding of how LSA works. LSA starts with a large corpus of meaningful texts, literally millions of words collected from books, newspaper articles, websites, and other sources. This semantic space is a mathematical representation of the semantic relationships among words based on their co-occurrences in the corpus (Hu et al., 2007).

Words that are frequently found just a few words away from other words in texts are represented as having similar meanings based on their
lexical co-occurrence. LSA represents each word as a vector of up to 500 dimensions, and the similarity between any two texts is computed using a mathematical formula. Dimensions with small values are removed, keeping only the most important dimensions corresponding to larger values (Hu et al., 2007).

I used the following illustration with my students. Consider the semantic relatedness of the words “peanut butter” and “jelly.” Peanut butter and jelly occur together frequently, and so would be represented as having similar semantic meaning. (One may consider this relationship blatant.) The similarity is expressed as a real number, typically between 0 and 1, with a higher number representing greater similarity. Now consider the words “peanut butter” and “bologna.” Presumably, peanut butter and bologna occur relatively rarely in the same sentence because most people don’t eat them together. However, these two words frequently co-occur with many of the same other words, such as “bread,” “sandwich,” “lunch,” and the like. Because of this latent similarity of lexical context, LSA also recognizes the similarity of “peanut butter” and “bologna.” Importantly, this relationship is based solely on lexical co-occurrence, rather than human judgment, and for the semantic processing engines used by SKO, the numeric vectors representing the meaning of words have already been established.

Because of the way the meaning of individual words and whole passages is represented mathematically, it is just as easy to use LSA to generate a measure of the similarity between two words, a word and a longer text passage, or two text passages (Hu et al., 2007). This is done by the same type of mathematical comparison of the representation of each word or passage as a numeric vector in the semantic space, where again, the semantic similarity is expressed as a real number between 0 and 1, with a higher number representing greater similarity. LSA has been shown to be successful in determining the semantic similarities of a range of texts (Hu et al., 2007). Thus, it is possible for SKO to compute the similarity between an expectations text created by the developer and text entered by a user. Significantly, the similarity index or coverage score (CO) is not dependent on having an exact word match. If working properly, synonyms and paraphrases will also receive a high score.

With SKO, LSA must be properly configured for the interactions to be successful. SKO requires expectations texts that reflect the gist of a good answer to the tutor’s questions so that it can compare input from participants to those expectations. If these expectations are not properly constructed, the tutor will be unable to make appropriate comparisons and, thus, be unable to respond appropriately.
SKO also needs a defined corpus of text that it can use to determine the mathematical similarity of texts. SKO is capable of using several such corpuses, but the most appropriate one must be identified. SKO also has many settings that can be adjusted to determine how it makes comparisons, such as the minimum association strength for words to be considered for comparisons. These settings must be calibrated as well, so that the tutor is best able to respond to participants’ answers. Finally, the actual responses of the tutor must be created so that they will best respond to the potential answers participants will create. The most accurate comparison between a participant’s responses and expectations would be wasted if the tutor could not use it to respond meaningfully.

Overall, the goal is to ensure that SKO is making appropriate determinations about the similarity between target texts and input entered by users. These target texts are referred to as expectation texts in SKO (Wolfe et al., 2012). Because semantic processing engines rely solely on lexical co-occurrence, the trick is to get the system to make comparisons based on the deeper meaning of the target expectations text (the gist; Wolfe, 2006, Wolfe et al., 2013) rather than on superficial characteristics (verbatim elements or their combinations). Thus, the numeric score generated by the comparisons will only be appropriate if the semantic engine has been provided with the proper expectations. If the expectations do not adequately represent the deep meaning of the text, the engine will not be able to make appropriate determinations about the similarity of relevant texts with similar deeper meanings, and [or about?] the dissimilarity of irrelevant texts with only surface similarities.

As part of his master’s research, Miami University psychology graduate student Colin Widmer created an Intelligent Tutoring System composed of four modules for teaching civics that he used in an experiment comparing the consequences of generating explanations, generating arguments, or no dialogue to a control group (Widmer, 2013). We used one of those modules, on the Electoral College, in this project for Learning & Cognition students. The module has a brief didactic presentation about the role of the Electoral College in American presidential politics, followed by a dialogue in which the avatar asks the learner to explain how the Electoral College works. Figure 1 is a screen shot from that tutorial. The avatar in the upper right corner has asked the learner to “Please explain how the Electoral College works to determine the President of the United States.” The learner first typed, “Each state gets a number of electoral college votes equal to its number og senators and representatives” [sic]. The bar on the left indicates that this input earned a coverage (or CO) score of 0.26. The avatar responded, “Great job keep going.” On the second turn the learner
added, “Thus, populous states such as California have relatively little influence while small states like Wyoming have relatively more influence.” The avatar did not respond, and the CO score rose to .348. Following this, the learner typed, “Electoral College” (not pictured in Figure 1). The CO score remained at 0.348, and the avatar responded, “We’ve been over this! Could you talk more about the role of electors.” Thus, the avatar was able to compare the expectations text to the learner’s input and respond appropriately—encouraging more elaboration when the learner was on track, and also recognizing when the learner was simply repeating a phrase from earlier and responding with an attempt to refocus the learner.

**Tutorial Dialogue Assignment**

Groups of three students were given a copy of the Electoral College tutorial with the expectations texts and the feedback rules removed. Following the technique developed by Wolfe and colleagues (2013), their task was to create and refine the expectations text, calibrate all of the settings, and create appropriate feedback rules and feedback. The assignment sheet given to the students is shown in Appendix A, and the scoring rubric used to evaluate the assignment, distributed when the assignment was given, is shown in Appendix B.

The general procedure was for students to go through the tutorial and write the essence of a good response to the question, then strip away small function words such as “of” and “the” that have too many associations to be useful. When students had an expectation text of fewer than 100 words, the next step was to feed back that identical text one sentence at a time, refining or tweaking the text and parameters until the final CO score was as high as possible. [How did they know how high was possible?]

The next task was to ensure that SKO can distinguish between relevant and irrelevant texts. This was accomplished by feeding in an irrelevant text, in this case the lyrics to “Take Me Out to the Ball Game,” making adjustments as necessary to make sure that this receives a low CO score compared to the original identical text. From here, the procedure outlined by Wolfe et al. (in press) [under review?] is to feed in more and less appropriate answers and make adjustments to ensure that SKO can distinguish between good and poor answers and that appropriate answers show successive improvements in CO score from one conversational turn (a sentence) to the next. The students then had to design appropriate feedback rules and feedback for more and less appropriate responses at various turns. We devoted three class periods to working on this assign-
ment, with graduate students Colin Widmer, Elizabeth Cedillos-Whynott, and me moving among groups and helping students with the assignment. These sessions were interspersed with classes devoted to discussions of relevant readings.

As can be seen in the scoring rubric (see Appendix B), groups that achieved basic functionality would earn a grade no lower than the C range, and those whose ITS could distinguish between the identical expectations text and the irrelevant text would at least earn a grade in the B range. These were tests that the students could run for themselves. To earn a grade in the A range, the tutor had to return a CO score of greater than 0.5 for a good answer provided by one of Widmer’s (2013) research participants that was not shared with the students as they worked on the assignment [This sentence is hard to follow]. In this way, we were able to simulate the task of creating an ITS for users who provide novel and perhaps unexpected answers. Pluses and minuses in scores were deter-
mined based on my professional judgments about whether the specific verbal feedback offered by the tutor would be likely to facilitate learning based on psychological concepts we had read and discussed.

The last class of the semester was devoted to trying out the tutorials created by other groups and then engaging in a seminar discussion of the experience. Before the end of the class period, I left the classroom, and, with the permission of the Internal Review Board, graduate student Audrey Weil administered an anonymous survey (see Appendix C) and asked those willing to stay after class to stay for a brief focus group discussion. Her notes from that session are provided in Appendix D.

**Results**

Next, we will examine student outcomes based on the evaluation of their work using the scoring rubric (see Appendix B), self-reported student learning results from the survey (see Appendix C), and student focus group comments (see Appendix D). All seven groups developed tutorial dialogues that could distinguish the expectation text they wrote from irrelevant text. Indeed, all but one group created tutors that were able to respond to good novel answers appropriately. Moreover, four of seven groups produced what I judged to be particularly helpful feedback based on psychological theory, and none produced unhelpful feedback. Thus, all of the groups met the criteria I established for successful projects. Interestingly, grades on this assignment correlated positively with grades on the rest of the course assignments ($R[21] = 0.25$)—despite a narrow range of grades and the fact that all group members received the same grade on this assignment. Thus, the extent to which students were successful with this project is associated with success on papers and other assignments.

Appendix C presents the mean and standard deviation for answers to 10 [Likert-style?] survey questions and an open-ended question from the 7 students who chose to answer it. “Compared to only reading and discussing articles about ITS, the ITS Dialogue Assignment helped me better understand the way ITS actually function” yielded a mean score of 6.38 on a 7-point scale, with 57% endorsing “strongly agree.” “Compared to only reading and discussing articles about ITS, the ITS Dialogue Assignment helped me better understand the strengths and weaknesses of ITS” yielded a mean score of 6.29, with 48% endorsing “strongly agree.” When asked, “If the professor teaches this course again, I recommend that he keeps the ITS Dialogue Assignment as part of this course” the mean was 6.05 and 45% endorsed “strongly agree.” By way of contrast, “Compared to only reading and discussing articles about ITS, the ITS Dialogue
Assignment was pretty much a waste of time” had a mean of 2.14, and “To be honest, I didn’t have the motivation to get much out of the ITS Dialogue Assignment” produced a mean of 2.55. [Note: I’m not sure I see the value of simply restating the results here that are in Appendix C. Am I missing something?]

The question of motivation produced a strong negative correlation with ratings of whether it was a good learning experience ($R_{20} = -0.78$). It also correlated negatively with [ratings of whether it?] helped me [students?] understand ITS ($R_{20} = -0.55$) and with ratings of whether it helped me [students?] understand the strengths and weaknesses of ITS ($R_{20} = -0.26$). However, the motivation item was positively correlated with the waste of time question ($R_{20} = 0.43$). These findings suggest that students by and large believed this to be a positive learning experience, and that much of the variance in negative outcomes could be accounted for by self-reported student motivation.

Students’ comments were generally quite positive (see Appendix D). Students not only used positive adjectives, but they were also able to identify specific benefits, such as the revelation that changing LSA modules from the general domain to the politics text corpus improved the performance of the system. In terms of criticisms, some students lamented the fact that they didn’t have more time for the assignment. There were also criticisms that the SKO technology is not “user friendly” for developers, particularly compared to commercial software. Finally, some students wished that they had more detailed instructions on how to create the expectations texts and calibrate SKO. Given the success achieved by all 7 groups, I argue it was a better learning experience for students to struggle with the assignment as a professional developer would (Wolfe et al., 2013) rather than spoon-feeding students “cookbook solutions.”

**Discussion**

The fact that small groups of undergraduate students without prerequisite technical knowledge were able to successfully modify an Intelligent Tutoring System to create pedagogical agents with the ability to engage in tutorial dialogues is a testament to SKO, in particular, and more generally to the rapid advances in discourse technologies. I share the widespread student perception that this was a successful learning experience. Indeed, given the pace at which these technologies are impacting our lives, it may be that some students some day will look back on this learning experience as important in helping them understand a technological landscape quite different from the one we inhabit today.
This assignment could be expanded to include an empirical research experience, in which students would not only create ITS, but also use them in psychology experiments to test hypotheses about human learning and cognition. An obvious weakness of the data reported here is that I did not conduct a randomized, controlled experiment. The best way to study the effectiveness of this experience would be to randomly assign students to either a tutorial dialogue or no tutorial dialogue condition, predicting changes in their knowledge, attitudes, and beliefs. Even if I had used a pre-test/post-test design, it would have been difficult to ascertain with certainty whether any effects were due to the exercise or reading and discussing articles about ITS. Indeed, I designed this exercise to complement students' engagement with the peer-reviewed journal literature rather than to replace it. Unfortunately, it would be difficult to conduct a randomized, controlled experiment of the sort outlined above. First, Learning & Cognition is not taught every year at Miami University, and I am not scheduled to teach it again. Second, the course typically draws about 25 students, which is inadequate to achieve adequate statistical power. More important, however, I have access to SKO only through a federal research grant (Wolfe et al., under review). Thus, although this assignment arguably exemplifies a healthy marriage between research and teaching, I may not have the opportunity to repeat this experience with other students in the future.

For many years, I have designed learning experiences to help students gain an intuitive feel for the gist for the phenomena they are investigating (Reyna, 2012), including statistical intuitions (Fisher & Wolfe, 2012), non-linear digital texts, (Wolfe, 1995), and educational courseware (Wolfe, 1992). Indeed, early in the technological revolution in higher education, my colleagues and I argued that this approach constitutes a “Miami Model” of Internet-intensive higher education (Wolfe et al., 1998). While financial pressures drive us to use technology to cut costs, I argue that we must also use emerging technologies to prepare students to be successful in a rapidly changing world.

Readers who have interacted with discourse technologies such as Siri know that emerging technologies are fallible—in ways that are sometimes even comical. Yet we are entering a new era of discourse technologies, and computer-based natural language processing is rapidly moving out of the laboratory and into everyday use. Professionals striving for excellence in college teaching should be seriously engaged with discourse technologies in our research and our teaching, if for no other reason than to ensure these systems are used to their full potential. Regardless of their chosen profession, it is not enough for students to be “super users,” confident in
their ability to find the nearest bar, but essentially ignorant of how their hand held device got them there. Today, a truly educated person must have a basic sense of the inner workings, an understanding of the “back end” of emerging discourse technologies.

References


Acknowledgments

The project described was supported by Award Number R21CA149796 from the National Cancer Institute. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Cancer Institute or the National Institutes of Health. We thank the National Cancer Institute for its support. We also wish to thank Xiangen Hu for allowing us to use AutoTutor Lite in this exercise.

Christopher R. Wolfe is a cognitive psychologist, past president of the Society for Computers in Psychology, and Professor of Psychology at Miami University. His research seeks an integrated understanding of higher-order thinking, conceptual learning, and the potential of information technologies. He teaches courses on learning and cognition, psychology of language and thought, and the psychology of medical decision-making.

Colin L. Widmer is a graduate student studying cognitive psychology with Christopher Wolfe. His research involves the use of discourse technologies with higher-order cognition and learning, as well as human-computer interactions involving natural language. He is interested in teaching courses on cognition and language. Audrey M. Weil is a graduate student in Christopher Wolfe’s lab at Miami University and is pursuing her degree in cognitive psychology. Her current research focuses on understanding the interactions between working memory capacity and reasoning strategies and how these two constructs work together to produce sound reasoning. She has taught a statistics and methods lab, and she hopes to one day teach cognitive and biopsychology classes at a small liberal arts college. Elizabeth M. Cedillos-Whynott earned an M.S. in Psychology at The University of Texas at San Antonio in 2010 and an M.A. in Psychology at Miami University in 2013. She is currently completing her Ph.D. at Miami University with a focus on the role of video games in influencing higher-order cognitive processes.
Appendix A

Intelligent Tutoring System Dialogue Assignment
(100 points small-group project)

Sharable Knowledge Objects (SKO) is a discourse technology developed by Dr. Xiangen Hu at the University of Memphis that uses Artificial Intelligence techniques. It allows human learners to interact with an animated intelligent agent in "natural language" (English). We have received special permission to use SKO in this course (and only for this purpose). SKO uses a technique called Latent Semantic Analysis (LSA) to "understand" what people type into a dialogue box and then respond appropriately to help them learn. Miami graduate student Colin Widmer developed an interactive tutorial on the Electoral College using SKO as part of his Master's thesis research. Your task, in small groups of 2-4, is to build on this work creating a brief interactive dialogue guided by psychological research on tutoring that will enable a learner to interact with the Electoral College tutorial in English. Specifically, you will keep the didactic portion of the Electoral College tutorial without making any alterations, and use all of the SKO tools at your disposal to make the program engage a learner in a tutorial dialogue (i.e., talk with the learner much as a human tutor would). No prerequisite technical knowledge is assumed, and we will devote some class time to this project.

Upon completion of this project, a person should be able to interact verbally with the tutor, and the tutor should encourage learners to engage in self-explanation and elaborate upon their answers. Thus, the tutor should respond appropriately to what the learner is saying. The tutor should be able to engage in a meaningful dialogue (that concludes with the instruction to click on the Finish button). The tutor should be able to distinguish between good and poor input from the learner and respond accordingly. For details of how this project will be evaluated, please consult the scoring rubric below.

All Tutorial Dialogue Groups receive the same grade unless problems arise. The tutorial dialogue should include content from the entire tutorial, beginning with the link below:

The last part of the tutorial that you edit can be found at http://think.psy.muohio.edu/home/DialogueGRowpLinks.htm. Each group will edit their own tutorial dialogue. When it is complete, let the world view it.
Appendix B
Scoring Rubric for
Intelligent Tutoring System Dialogue Assignment
(100 points small-group project)

Note: This rubric climbs a ladder from basic functionality to responding to novel good answers appropriately. Projects will be awarded the highest grade earned (i.e., the highest level they achieve). Pluses and minuses (+/-) will be assigned separately, as described below.

0. Lacks Basic Functionality. The tutor [Should it be “tutor” or “tutorial”?] is not able to respond appropriately to users and tell them to click on the Finish button at the appropriate time. Grade: F or D range at the discretion of the professor.

I. Basic Functionality. The tutor is able to respond to learners and tell them to click on the Finish button at the appropriate time. Grade: At Least C Range.

II. Distinguishes Expectation Text Itself From Irrelevant Text. When the expectation text you developed is fed into the tutor as input one sentence at a time it earns a final coverage (CO) score of at least 0.50. When irrelevant text (“Take Me Out to the Ball Game”) is fed into the tutor one sentence at a time it earns a final (coverage) CO score much less than the identical expectation text. Grade: At Least B Range.

III. Responds to Novel Good Answers Appropriately. When a good answer of several sentences (that you have not seen before) is fed into tutor as input one sentence at a time it earns a final coverage (CO) score of at least 0.50. Grade: A Range.

Pluses and Minuses (+/-). In the judgment of the professor, the specific verbal feedback offered by the tutor would be likely to facilitate learning based on psychological concepts. For example, it encourages the learner to engage in self-explanation and provide a more elaborate appropriate response. Pluses (+) are earned for particularly helpful feedback, and minuses (-) would be earned for unhelpful feedback.

Examples. A tutorial that distinguishes expectation text itself from irrelevant text and provides feedback that is not particularly helpful or unhelpful would earn a score of 85/100. A tutorial that responds to novel good answers appropriately, but tells the learner to “Go to Hell” when a poor answer is provided, might earn a grade of 90/100 (an A-). Although the tutorial warrants the A range for novel good answers, the inappropriate feedback would bring the score down. Finally, a tutorial that achieves only basic functionality but would have given very helpful feedback if only it had worked a little better might earn a Grade of 79/100, corresponding to a C+. 
Feedback on the Intelligent Tutoring System Dialogue Assignment

The purpose of this research is to help your professor improve teaching practices. A graduate student will administer this survey, and your professor will not look at these data until after the end of the semester. You may skip any questions or refuse to participate without any penalty. Please respond to each question about the assignment creating [to create an?] Intelligent Tutoring System (ITS) Dialogue. For each question, compare the experience of engaging in this assignment to simply reading and discussing articles about intelligent tutoring systems. Please circle the appropriate number using a scale from 1 (strongly disagree) to 7 (strongly agree).

1. Compared to only reading and discussing articles about ITS, the ITS Dialogue Assignment helped me better understand the way ITS actually function. **Mean = 6.38 (SD = 0.86)**

2. Compared to only reading and discussing articles about ITS, the ITS Dialogue Assignment helped me better understand the strengths and weaknesses of ITS. **Mean = 6.29 (SD = 0.85)**

3. Compared to only reading and discussing articles about ITS, the ITS Dialogue Assignment was pretty much a waste of time. **Mean = 2.14 (SD = 0.79)**

4. Compared to only reading and discussing articles about ITS, the ITS Dialogue Assignment helped me better understand the psychological literature on learning from ITS. **Mean = 5.19 (SD = 1.08)**

5. Compared to only reading and discussing articles about ITS, the ITS Dialogue Assignment helped me apply key psychological concepts to designing learning environments. **Mean = 5.29 (SD = 1.19)**

6. To be honest, I didn’t have the motivation to get much out of the ITS Dialogue Assignment. **Mean = 2.55 (SD = 1.39)**

7. I am glad that the ITS Dialogue Assignment was part of this course. **Mean = 5.85 (SD = 0.81)**

8. If the professor teaches this course again, I recommend that he keep the ITS Dialogue Assignment as part of this course. **Mean = 6.05 (SD = 1.10)**

9. The criteria for evaluating the ITS Dialogue Assignment was clear, fair, and understandable. **Mean = 5.90 (SD = 0.79)**

10. The ITS Dialogue Assignment was a good learning experience. **Mean = 5.95 (SD = 0.94)**

11. Please write any comments you have about the ITS Dialogue Assignment:
Appendix C (continued)

Feedback on the Intelligent Tutoring System Dialogue Assignment

Responses:
I enjoyed it, personally. Think it’d be good to bring back.
I really enjoyed this unit! I would recommend that it is taught again.
Loved it!
Maybe more time to tweak it?
Really cool to work with this new technology!
Have in depth instructions regarding the development and refinement of "good" expectations text would be good.
I know there is not much about the appearance of the ITS, but it’s really not user friendly; this makes it much more difficult to use. It’s a nice break from a paper, though.