Human or Computer? AutoTutor in a Bystander Turing Test

Natalie Person[1], Arthur C. Graesser[2], & The Tutoring Research Group[2]

[1] Rhodes College, Department of Psychology, Memphis, TN 38112
person@rhodes.edu
[2] University of Memphis, Department of Psychology, Memphis, TN 38152-3230
a-graesser@memphis.edu

1 Introduction

Since the development of the first digital computer in the 1940s, the notion of computer intelligence has received considerable attention from computer scientists, philosophers, and psychologists. The question of whether it is possible to create a computer program that possesses human intelligence has spurred much debate. Turing (1950) argued that computers are not capable of thinking and provided several theological, psychological, and sociological arguments in support of his position. To determine a computer program's intelligence, Turing proposed several benchmark methods. One such method requires humans to decide whether they are interacting with an actual computer program or another human via computer mediation. According to Turing, a computer could be described as intelligent if it could deceive a human into believing that it was human. The two studies presented here were designed to determine whether AutoTutor could pass a variation of the Turing test, the Bystander Turing Test. The subsequent sections of this paper address the following: (1) the AutoTutor system, (2) the Bystander Turing Test, (3) the two empirical studies, and (4) the conclusions of the studies.

2 Description of AutoTutor

AutoTutor is a generic computer tutor architecture that can be used for a variety of content domains [3], [5], [10], [12]. The Tutoring Research Group (TRG) has recently developed two versions of AutoTutor, one for computer literacy and one for conceptual physics. The computer literacy AutoTutor is designed to help students learn basic computer literacy topics covered in an introductory course (e.g., hardware, operating systems, and the Internet). The conceptual physics AutoTutor is designed to help students learn Newtonian physics.

AutoTutor's architecture is comprised of six major modules: (1) an animated agent, (2) a curriculum script, (3) language analyzers, (4) latent semantic analysis (LSA), (5) a dialog move generator, and (6) a Dialog Advancer Network [7], [9], [11], [12], [13], [15], [16]. AutoTutor initiates the conversation with the learner by selecting a question or problem from the curriculum script for the learner to solve. Students learn about computer literacy or physics by engaging in a conversation with the animated

S.A. Cerri, G. Gouardères, and F. Paraguaçu (Eds.): ITS 2002, LNCS 2363, pp. 821–830, 2002.

agent. AutoTutor scaffolds the conversation with a series of dialog moves that are frequently used by effective human tutors [4], [8]. The dialog moves included in AutoTutor are Pump, Prompt, Hint, Assertion, Correction, Summary, and three kinds of Short Feedback (positive, negative, and neutral).

Although the selection of dialog moves is complex in AutoTutor, a summary of this process is as follows. After each typed student contribution, a series of language analyzers operate on the words in the student's contribution so that the contribution can be classified into one of five speech act categories: Assertion, WH-question, Yes/No question, Frozen Expression, or Prompt Completion. The quality of the Assertion classifications is determined by latent semantic analysis (LSA). LSA is also used to monitor several other parameters that are important in tutoring (e.g., topic coverage, student ability). The dialog move generator is controlled by a series of production rules that utilize the LSA parametric data. The dialog move generator selects one or a combination of pedagogically appropriate dialog moves from the curriculum script. These moves are conveyed to the student via the animated agent. The Dialog Advancer Network (DAN) manages the turn-taking and provides AutoTutor responses to all of the speech act categories.

3 The Bystander Turing Test

The original Turing test is based on the Imitation Game. In the Imitation Game, the participants are a man, a woman, and an interrogator. The interrogator is physically separated from the man and woman. The object of the game is for the interrogator to decipher which participant is male and which is female by evaluating their responses to questions posed by the interrogator. Turing proposed replacing one of the humans (i.e., the man or woman) with a computer. If the interrogator cannot discern whether the responses are generated by the computer or by the human, the computer program is said to pass the Turing Test. Specifically, the computer program is emulating human thought and intelligence if it can process human language, utilize knowledge it has been given before and receives during the interactive session, and use stored knowledge to engage in an intellectual conversation with the learner [14].

The Bystander Turing Test (BTT) is a variation of the original Turing Test. In the BTT, participants rate whether particular dialog moves in tutoring transcripts are generated by AutoTutor or by skilled human tutors. We converged on this variation of the original Turing Test for two reasons that are based on our observations of human tutors attempting to tutor students in computer-mediated environments. First, human tutors have tremendous difficulty responding in real-time to student input. Given that AutoTutor responds immediately to student contributions, noticeable time lags caused by human tutors' need to think would potentially influence an interrogator's decisions. Second, human tutors frequently make errors when typing their responses. Hence, AutoTutor's text-to-speech generator would mispronounce a considerable number of words, and unquestionably, affect an interrogator's judgments.

The conversations included in both BTT studies were constructed in the following way. Two hundred and eighty-two conversations were randomly selected from several thousand conversations that had taken place between college students and the computer literacy AutoTutor. In each of the 282 conversations, a particular AutoTutor dialog move was deleted along with all subsequent conversational turns. Six skilled human computer literacy tutors were asked to read each conversation and then fill in the blank lines with what they would say to that student at that juncture in the conversation.

After the human tutor responses (i.e. dialog moves) were collected, packets containing 36 conversations were assembled. Each packet contained 18 conversations in which all tutor dialog moves were generated by AutoTutor and 18 conversations in which the last dialog move in the conversation was generated by a human tutor. To the extent possible, the AutoTutor dialog move categories (e.g., Pump, Prompt, Assertion) were evenly distributed across the 18 authentic AutoTutor conversations. Participants in each study were asked to read the entire conversation and evaluate the last dialog move. The evaluation questions differed in the two studies and will be discussed in the next two sections.

3.1 Study 1

The participants were 64 students who were either enrolled in a computer literacy course or who had completed the course at the University of Memphis. The participants received extra credit in a course for their participation. All participants read and signed an informed consent form before the experiment began. Testing packets were then given to the participants which included 36 conversations, 18 of the conversations ended in an AutoTutor dialog move and 18 ended in a human tutor dialog move. After each conversation, participants rated the last tutor dialog move on a six-point scale, 1 indicating the dialog move was definitely generated by a human and 6 indicating the dialog move was definitely generated by a computer (see Table 1). After providing ratings for all 36 conversations, participants were asked to describe any strategies they had developed for identifying which dialog moves were generated by the computer and which were generated by humans. The analysis of these open-ended responses will not be discussed in this paper.

3.2 Study 2

The participants were 24 undergraduates at Rhodes College, a small liberal arts college located in Memphis, Tennessee. The students participated in the study to fulfill research requirements for introductory psychology classes. The materials and procedure for Study 2 differed from Study 1 in the following ways. After reading each conversation, participants provided three six-point scale ratings for the last tutor dialog move in each conversation. The additional questions were added to assess the teaching effectiveness and conversational appropriateness of the dialog moves. The three rating questions used in Study 2 are provided in Table 1. After the participants completed all 36 dialog move ratings, they were asked to describe any strategies they had developed for identifying the speakers of the dialog moves. Lastly, participants were asked to complete a computer knowledge questionnaire. This questionnaire was included to determine whether computer knowledge is a moderating factor in participants' ability to discriminate human versus computer tutors.

Table 1. Dialog Move Assessment Questions Used in Study 1 and 2

(Used in Study 1 and 2)

1. The last tutor response/question is

1	2	3	4	5	6
Definitely Human	Probably Human	Not sure, but guess Human	Not sure, but guess Computer	Probably Computer	Definitely Computer

(Study 2 only)

2. In terms of conversational appropriateness, the last tutor response/question is (do not circle between the bars):

1	2	3	4	5	6
Definitely Inappropriate	Probably Inappropriate	Not sure, but guess Inappropriate	Not sure, but guess Appropriate	Probably Appropriate	Definitely Appropriate

(Study 2 only)

3. In terms of effective teaching, the last tutor turn response/question is

1	2	3	4	5	6
Definitely Ineffective	Probably ineffective	Not sure, but guess Ineffective	Not sure, but guess Effective	Probably Effective	Definitely Effective

4 Results and Discussion

A series of independent samples t-tests were performed to determine whether the human and computer means differed for the three six-point scale assessment questions used in Study 1 and 2. The means for the three questions are reported in Table 2. A series of one-way ANOVAs were performed to determine whether participants were relying on dialog move category information when providing ratings for the three assessment questions. Means and likelihood values from these analyses are reported in Tables 3, 4, and 5. The results from all of these analyses are reported and discussed below.

4.1 Who Said It? (Question 1)

Independent samples t-tests were performed to determine whether participants could distinguish AutoTutor versus human-generated dialog moves in Study 1 and Study 2. For the "Who said it?" question, mean ratings on the six-point scale for the last dialog move in each conversation were compared (see Table 2). Results from both studies indicated that participants could not discriminate between dialog moves generated by AutoTutor and those generated by humans, *t*(142) = 1.45, p > .10; *t*(142) = .68, p > .10; Study 1 and Study 2, respectively. Hence, AutoTutor passed the Bystander Turing Test when observations were collected at a fine-grained level (i.e., at the level of individual dialog moves, as opposed to interacting with AutoTutor versus a human for 30 minutes). Simply put, the college students could not tell whether the dialog move was generated by a computer or by a human computer literacy tutor. AutoTutor therefore does a fairly good job of simulating skilled human tutor dialog moves.

Table 2. Means and Standard Deviations for Assessment Questions in Study 1 and Study 2

Study 1

Question	Computer			Human		
	Mean	s.d.	n	Mean	s.d.	n
1. Who said it?	3.52	0.44	72	3.62	0.38	72

Study 2

Question	Computer			Human		
	Mean	s.d.	n	Mean	s.d.	n
1. Who said it?	3.72	0.91	72	3.82	0.76	72
2. Conversationally appropriate?	4.07	0.82	72	4.13	0.77	72
3. Effective teaching?	3.76	0.78	72	3.81	0.67	72

One alternative explanation of why the speaker discrimination was so low involves the dialog move categories. Perhaps the college students used the dialog move category information to diagnostically decide whether a computer or human produced the dialog move. College students may have been penalized in their predictions if they were poor guessers for some categories, but accurate guessers for others. We do know that the distribution of dialog moves of human tutors is radically different from the distribution of dialog moves in the AutoTutor sample. For example, human tutors tend to generate helpful Assertions whereas AutoTutor tries to get the learner to do the talking through Hints and Prompts.

Although sophisticated guessing could conceivably account for the poor discrimination in college students predicting computer versus human, a closer look at the data would reject this alternative explanation. We discovered that the likelihood that the college students believed a dialog move was generated by a computer was absolutely equivalent for all six dialog move categories included in Study 1 and all seven dialog move categories included in Study 2. The likelihood values were computed by considering the proportion of four, five, and six ratings on the "Who Said It?" question for each dialog move category (see Table 1). The means and likelihood values of the dialog move categories are reported in Table 3. One-way ANOVAs indicated that there were no significant differences among the dialog move categories. Thus, the college students did not show a bias among the different dialog move categories by rating particular categories as more likely to be generated by a computer.

Another potential explanation for low speaker discrimination is that the human tutors adopted the conversational style of AutoTutor. Although the human tutors were instructed to generate exactly what they would say to a student at a particular point in a tutoring conversation, the human tutors may have been inadvertently affected by the previous AutoTutor dialog moves. Thus, the human tutors may have generated succinct dialog moves that are inherent in AutoTutor's curriculum script.

Table 3. Means and Likelihood Values of Dialog Move Categories for the "Who Said It?" Question

Study 1

Dialog Move Category	Mean	s.d	Likelihood	n
Pump	3.47	0.43	0.50	24
Prompt	3.66	0.41	0.56	22
Prompt completion	3.55	0.48	0.50	26
Hint	3.68	0.36	0.55	20
Assertion	3.49	0.37	0.51	29
Correction	3.60	0.40	0.53	23
Total	3.57	0.41	0.52	144

Study 2

Dialog Move Category	Mean	s.d.	Likelihood	n
Pump	3.93	0.78	0.60	27
Prompt	3.72	0.93	0.59	27
Hint	4.06	0.93	0.64	21
Assertion	3.62	0.83	0.55	38
Correction	3.67	0.82	0.56	21
Summary	3.50	0.53	0.58	6
Feedback	3.94	0.01	0.72	3
Total	3.77	0.84	0.59	143

Study 1: One-way ANOVA for Means, $F(5,138) = 1.07$, $p > .10$; One-way ANOVA for Likelihood Values, $F(5,138) = 1.20$, $p > .30$

Study 2: One-way ANOVA for Means, $F(6, 136) = 0.94$, $p > .10$; One-way ANOVA for Likelihood Values, $F(6,136) = 0.50$, $p > .10$

(NOTE: The Prompt Completion Category was not included in the Study 2 analyses because there was only one occurrence.)

4.2 Conversationally Appropriate? (Question 2)

For the question designed to assess conversational appropriateness (Question 2), an independent-samples t-test yielded no mean differences between human and computer generated dialog moves, $t(142) = 0.44$, p > .50. That is, participants considered dialog moves generated by AutoTutor to be as conversationally appropriate as those generated by human tutors (AutoTutor mean = 4.07, Human mean = 4.13). The fact that the means did not differ and were closer to the more favorable extreme of the six-point scale is a promising indicator that AutoTutor is a competent conversational partner. As mentioned in the discussion of the "Who said it?" (Question 1) results, the human tutors may have adopted the conversational style of AutoTutor rather than generating the dialog moves they would have produced in an actual human-to-human tutoring session. Even if this was the case, the participants in Study 2 considered the dialog moves in both speaker conditions to be conversationally appropriate. Hence, even if AutoTutor does not emulate the style of human tutors in human-to-human sessions, AutoTutor does deliver dialog moves that were deemed acceptable and believable.

One of the fundamental design goals of AutoTutor is to create a system that helps students construct answers and explanations by delivering dialog moves that are pedagogically effective and conversationally appropriate. In order to determine whether particular dialog move categories were considered more conversational than others and to determine whether some dialog move categories were too computer-like, we compared the means and likelihood values of the dialog categories in two one-way ANOVAs. The means, likelihood values, and statistical results for the "Conversationally Appropriate?" question are reported in Table 4. Recall the means are the average six-point scale ratings on the "Conversationally Appropriate?" question (see Table 1), and the likelihood values are the proportions of four, five, and six ratings for each dialog move category. For example, 60% of the Pump moves received a four, five, or six rating irrespective of the speaker. The results from the one-way ANOVAs indicated no significant differences among the means or likelihood values for any of the dialog move categories. Thus, no dialog move category was considered more or less conversationally appropriate than any other, and all of the categories received high ratings approximately 70% of the time.

4.3 Effective Teaching? (Question 3)

The third assessment question addressed the pedagogical effectiveness of particular dialog moves. For the "Effective Teaching?" question, an independent-samples t-test indicated no mean differences between computer- and human-generated dialog moves, *t*(142) = 0.30, p > .10 (see Table 2). Therefore, participants were not using speaker characteristics when making holistic judgments about the pedagogical quality of the dialog moves.

Table 4. Study 2 Means and Likelihood Values of Dialog Move Categories for the "Conversationally Appropriate?“ Question

Dialog Move Category	Mean	s.d.	Likelihood	n
Pump	3.97	0.86	0.66	27
Prompt	3.94	0.86	0.64	27
Hint	3.84	0.88	0.62	21
Assertion	4.20	0.68	0.71	38
Correction	4.26	0.72	0.74	21
Summary	4.61	0.39	0.89	6
Feedback	4.67	0.44	0.89	3
Total	4.09	0.79	0.69	143

Study 2: One-way ANOVA for Means, *F*(6,136) = 1.65, *p* > .10; One-way ANOVA for Likelihood Values, *F*(6,136) = 1.52, *p* > .10

In keeping with the dialog move category analyses performed for the other assessment questions, we wanted to discern whether the participants viewed some dialog move categories as more effective teaching strategies than others. Two one-way ANOVAs were performed to determine whether statistical differences occurred among the means and likelihood values for the dialog move categories. Both one-way ANOVAs indicated significant mean and likelihood value differences among the dialog move categories. LSD post-hoc tests were performed to determine the specific differences among the means and likelihood values. All means, likelihood values, and statistical results for the "Effective Teaching?" dialog move analysis are reported in Table 5.

In the "Effective Teaching?" means analysis, participants considered Assertions, Corrections, Summaries, and Feedback moves to be more effective teaching categories than Pumps, Prompts, and Hints, p < .05. The likelihood analysis produced a similar pattern of results; Assertions, Summaries, and Feedback moves had significantly greater proportions of high ratings (participants rated them a 4, 5, or 6) than Pumps, Prompts, Hints, and Corrections, p < .05. The differences between particular dialog move categories for the "Effective Teaching?" question should take few educational researchers or practitioners by surprise. Assertions, Summaries, and Corrections, are information delivery moves that have different pedagogical functions; however, these moves do not require students to actively elaborate their own knowledge. Thus, participants showed clear preferences for dialog moves that minimized the cognitive effort of the students and maximized the informational output of tutors.

Table 5. Study 2 Means and Likelihood Values of Dialog Move Categories for the "Effective Teaching?“ Question

Dialog Move Category	Mean	s.d.	Likelihood	n
Pump	3.54^{a}	0.64	0.57^{a}	27
Prompt	3.53^{a}	0.78	0.51^{a}	27
Hint	3.52^{a}	0.90	0.55^{a}	21
Assertion	4.04^{b}	0.61	0.68^{b}	38
Correction	3.96^{b}	0.61	0.63^{a}	21
Summary	4.39^{b}	0.39	0.78^{b}	6
Feedback	4.33^{b}	0.17	0.89^{b}	3
Total	3.78	0.73	0.61	143

$^{a\ b}$ Means and likelihood values that share a superscript do not statistically differ.

Study 2: One-way ANOVA for Means, $F(6,136) = 3.91$, $p < .001$; One-way ANOVA for Likelihood Values, $F(6,136) = 2.82$, $p < .05$

These results are consistent with many educators's beliefs that students prefer to be passive recipients of information that is spoon-fed to them rather than learners who actively engage in the learning process by generating explanations and asking questions. Participants did consider the Feedback dialog moves to be effective teaching strategies; however, this finding should be interpreted with caution given the few instances of Feedback moves (n = 3) in the Study 2 sample.

5 Conclusions

A number of conclusions can be drawn from the findings of these two bystander Turing studies. The participants in both studies were unable to discriminate dialog moves that were generated by humans from those generated by AutoTutor. Therefore, AutoTutor is, to some extent, achieving many of the design goals of AutoTutor's developers by simulating the dialog moves of effective human tutors and delivering them in conversationally appropriate ways. Future studies, however, should include conditions in which randomly generated dialog moves from human tutors and AutoTutor are inserted in the turn that participants evaluate. If participants can discriminate random moves from non-random ones but cannot discriminate human versus computer, then we will have more compelling evidence that AutoTutor's dialog move selections are on par with those of human tutors.

The dialog move category analyses indicated that participants did not associate particular dialog categories with the speaker categories nor did they consider any of the categories to be more conversationally appropriate than others. However, the participants did consider some dialog move categories to be more effective teaching strategies than others. This finding reflects participants' beliefs about what constitutes effective teaching strategies as well as their beliefs about how students learn. Unfortunately, these beliefs are not compatible with current research devoted to effective pedagogy. This incompatibility between student beliefs and tutors' pedagogical goals presents the same problems for developers of intelligent tutoring systems as those frequently encountered by teachers and human tutors. Effective teachers and tutors encourage students to take an active role in the learning process by structuring tasks and dialogs in ways that force students to engage in behaviors such as asking questions, recognizing misconceptions, generating explanations, and synthesizing information from multiple sources [1], [2], [4], [13]. Such student behaviors are not only rare without teacher intervention but frequently cause cognitive distress in students. ITS developers must therefore strive to design systems that preserve effective pedagogy without overly frustrating students.

Acknowledgements. This research was supported by grants from the National Science Foundation (SBR 9720314 and REC 0106965) and the Department of Defense Multidisciplinary University Research Initiative (MURI) administered by the Office of Naval Research under grant N00014-00-1-0600. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of ONR or NSF.

References

1. Chi, M.T.H.: Constructing self explanations and scaffolded explanations in tutoring. *Applied Cognitive Psycholog, 10* (1996) S33-S49

2. Chi, M.T.H., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R.: Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Scienc, 13* (1989) 145-182
3. Graesser, A.C., Hu, X., Susarla, S., Harter, D., Person, N.K., Louwerse, M., Olde, B., & the Tutoring Research Group: AutoTutor: An Intelligent Tutor and Conversational Tutoring Scaffold. *Proceedings for the 10th International Conference of Artificial Intelligence in Education* San Antonio, TX (2001) 47-49
4. Graesser, A.C., Person, N.K., & Magliano, J.P.: Collaborative dialogue patterns in naturalistic one-to-one tutoring sessions. *Applied Cognitive Psychology 9* (1995) 1-28
5. Graesser, A.C., Person, N.K., Harter, D., & the Tutoring Research Group: Tactics in tutoring in AutoTutor. In the *ITS 2000 Proceedings of the Workshop on Modeling Human Teaching Tactics and Strategies* Montreal, Canada (2000) 49-57
6. Graesser, A.C., Wiemer-Hastings, K., Wiemer-Hastings, P., Kreuz, R., & TRG: AutoTutor: A simulation of a human tutor. *Journal of Cognitive Systems Research 1*(1999) 35-51
7. Graesser, A.C., Wiemer-Hastings, P., Wiemer-Hastings, K., Harter, D., Person, N., & the Tutoring Research Group: Using latent semantic analysis to evaluate the contributions of students in AutoTutor. *Interactive Learning Environments 8* (2000) 129-148
8. Person, N.K., & Graesser, A.C.: Evolution of discourse in cross-age tutoring. In A. M.O'Donnell and A. King (Eds.), *Cognitive perspectives on peer learning* Mahwah, NJ: Erlbaum (1999) 69-86
9. Person, N.K., Graesser, A.C., & the Tutoring Research Group: Designing AutoTutor to be an effective conversational partner. In the *Proceedings for the 4th International Conference of the Learning Sciences* Ann Arbor, MI (2000) 246-253
10. Person, N.K., Graesser, A.C., Bautista, L., Mathews, E.C., & the Tutoring Research Group: Evaluating Student Learning Gains in Two Versions of AutoTutor. In J.D. Moore, C.L. Redfield, & W.L. Johnson (Eds.) *Artificial intelligence in education: AI-ED in the wired and wireless future* Amsterdam, IOS Press (2001) 286-293
11. Person, N.K., Graesser, A.C., Harter, D., Mathews, E.C., & the Tutoring Research Group: Dialog move generation and conversation management in AutoTutor. *Proceedings of the AAAI Fall Symposium: Building Dialogue Systems for Tutorial Applications* Falmouth, MA: AAAI Press (2000) 45-51
12. Person, N.K., Graesser, A.C., Kreuz, R.J., Pomeroy, V., & the Tutoring Research Group: Simulating human tutor dialog moves in AutoTutor. *International Journal of Artificial Intelligence in Education 12* (2001b) 23-29
13. Person, N.K., Klettke, B., Link, K., Kreuz, R.J., & the Tutoring Research Group: The integration of affective responses into AutoTutor. *Proceedings of the International Workshop on Affect in Interactions* Siena, Italy (1999) 167-178
14. Turing, A.M.: Computing Machinery and Intelligence. In E.A. Feigenbaum & J. Feldman (Eds.) *Computers and thought*. New York: McGraw-Hill (1950)
15. Wiemer-Hastings, P., Graesser, A.C., Harter, D., and the Tutoring Research Group: The foundations and architecture of AutoTutor. *Proceedings of the 4th International Conference on Intelligent Tutoring Systems* Berlin, Germany: Springer-Verlag (1998) 334-343
16. Wiemer-Hastings, P., Wiemer-Hastings, K., and Graesser, A.: Improving an intelligent tutor's comprehension of students with Latent Semantic Analysis. *Artificial Intelligence in Education* Amsterdam: IOS Press (1999) 535-542