Why/AutoTutor: A Test of Learning Gains from a Physics Tutor with Natural Language Dialog

Graesser, A.C.[1], Jackson, G.T. [1]., Mathews, E.C.[1], Mitchell, H.H.[1], Olney, A.[1], Ventura, M.[1], Chipman, P.[1], Franceschetti, D.[1], Hu, X.[1], Louwerse, M.M.[1], Person, N.K.[2], and the Tutoring Research Group[1]

[1] Institute for Intelligent Systems
University of Memphis
Memphis, TN 38152

[2] Rhodes College
2000 N Parkway
Memphis, TN 38112

Abstract

Why/AutoTutor is a tutoring system that helps students construct answers to qualitative physics problems by holding a conversation in natural language. Why/AutoTutor provides feedback to the student on what the student types in (positive, neutral, negative feedback), pumps the student for more information, prompts the student to fill in missing words, gives hints, fills in missing information with assertions, identifies and corrects bad answers and misconceptions, answers students' questions, and summarizes answers. In essence, constructivist learning is implemented in a mixed-initiative dialog. Why/AutoTutor delivers its dialog moves with an animated conversational agent whereas students type in their answers via keyboard. We conducted an experiment that compared Why/AutoTutor with two control conditions (Read textbook, nothing) in assessments of learning gains. The tutoring system performed significantly better than the two control conditions on a test similar to the Force Concept Inventory.

AutoTutor and Why/AutoTutor

Why/AutoTutor is the most recent tutoring system in the AutoTutor series developed by the Tutoring Research Group at the University of Memphis. Why/AutoTutor was specifically designed to help college students learn Newtonian qualitative physics (Graesser, VanLehn, Rose, Jordan, & Harter, 2001), whereas the previous AutoTutor systems were on topics of introductory computer literacy (Graesser, Person, Harter, & TRG, 2001; Graesser, P. Wiemer-Hastings, K. Wiemer Hastings, & Kreuz, 1999) and military tactical reasoning (Ryder, Graesser, McNamara, Karnavat, & Pop, 2002).

This design of AutoTutor was inspired by explanation-based constructivist theories of learning (Aleven & Koedinger, 2002; Chi, deLeeuw, Chiu, LaVancher, 1994; VanLehn, Jones, & Chi, 1992) and by previous empirical research that has documented the collaborative constructive activities that routinely occur during human tutoring (Chi, Siler, Jeong, Yamauchi, & Hausmann, 2001; Fox, 1993; Graesser, Person, & Magliano, 1995; Moore, 1995; Shah, Evens, Michael, & Rovick, 2002). The process of actively constructing explanations and elaborations of the learning material allegedly produces better learning than merely presenting information to students. This is where human tutors excel in scaffolding learning, because they guide the students in productive constructive processes and simultaneously respond to the student's information needs.

Surprisingly, the dialog moves of most human tutors are not particularly sophisticated from the standpoint of today's pedagogical theories and those theories implemented in intelligent tutoring systems (Graesser et al., 1995). Human tutors normally coach the student in filling in missing pieces of information in an expected answer and they fix bugs and misconceptions that are manifested by the student during the tutorial dialog. Human tutors rarely implement *bona fide* Socratic tutoring strategies, modeling-scaffolding-fading, and other intelligent pedagogical techniques (Collins, Brown, & Newman, 1989). The argument has been made that it is the conversational properties of human tutorial dialog, not sophisticated tutoring tactics, that explain why human tutors facilitate learning (Graesser et al., 1995). Why/AutoTutor was designed to simulate the dialog moves of human tutors who coach students in constructing explanations.

Why/AutoTutor helps students learn by presenting challenging problems (or questions) from a curriculum script and engaging in mixed initiative dialog while constructing an answer. An example question is "Suppose a boy is in a free-falling elevator and he holds his keys motionless right in front of his face and then lets go. What will happen to the keys? Explain why." Such questions are designed to require about a paragraph of information (3-7 sentences) to answer. However, initial answers to these questions are typically only 1 or 2 sentences in length, even though students have more knowledge that is relevant to an answer. This is where tutorial dialog is particularly helpful. AutoTutor engages the student in a mixed initiative dialog that assists in the evolution of an improved answer and that draws out more of what the students know. AutoTutor provides feedback to the student on what the student types in (positive, neutral, negative feedback), pumps the student for more information ("What else?"), prompts the student to fill in missing words, gives hints, fills in missing information with assertions, identifies and corrects erroneous ideas and misconceptions, answers the student's questions, and summarizes answers. A full answer to the question is eventually constructed during this dialog.

Figure 1: Interface of Why/AutoTutor

Figure 1 shows the interface of Why/AutoTutor. The major question is selected and presented in the top-right window. This question remains at the top of the web page until it is finished being answered by a multi-turn dialog between the learner and Why/AutoTutor. The students use the bottom-right window to type in their contributions for each turn, with the content of both tutor and student turns being reflected in the bottom-left window. The animated conversational agent resides in the upper-left area. The agent uses either an AT&T or a Microsoft Agent speech engine to speak the content of AutoTutor's turns during the process of answering the presented question, dependent on the computational reseources the user has available.

The computational architecture of Why/AutoTutor and earlier versions of AutoTutor have been discussed extensively in previous publications (Graesser, Person et al., 2001; Graesser, VanLehn, et al., 2001; Graesser, Wiemer-Hastings et al., 2001), so this paper will provide only a brief sketch of the components. Why/AutoTutor was written in Java and resides on a Pentium-based server platform to be delivered across the web. The software residing on the server has a set of permanent databases that do not get updated throughout the course of tutoring. These include (a) the curriculum script repository consisting of questions, answers, and associated dialog moves, (b) lexicons, syntactic parsers, and other computational linguistics modules, (c) a question answering facility, (d) a corpus of documents, including a text book on conceptual physics, and (e) latent semantic analysis (LSA) vectors for words, curriculum content, and the document corpus. Why/AutoTutor uses LSA as the backbone for representing world knowledge about conceptual physics, or any other subject matter that is tutored (Olde, Franceschetti, Karnavat, Graesser, & TRG, 2002). LSA is a high-dimensional, statistical technique that, among other things, measures the conceptual similarity of any two pieces of text, such as a word, sentence, paragraph, or lengthier document (Foltz, Gilliam, & Kendall, 2000; Kintsch, 1998; Landauer, Foltz, & Laham, 1998). In Why/AutoTutor we use LSA to perform conceptual pattern matching operations when we compare student contributions to expected good answers and to misconceptions.

In addition to the modules mentioned above, Why/AutoTutor has a set of processing modules and dynamic storage units that maintain qualitative content and quantitative parameters. These storage registers are frequently updated as the tutoring process proceeds. For example, Why/AutoTutor keeps track of student ability (as evaluated by LSA from student Assertions), student initiative (such as the incidence of student questions), student verbosity (number of words per turn), and the progress in having a question answered by virtue of the dialog history. The dialog management module of AutoTutor flexibly adapts to the student by virtue of these parameters, so it is extremely unlikely that two conversations with AutoTutor are ever the same. The dialog

management module has an augmented finite state network and a special algorithm for selecting dialog moves to help fill in missing information in an ideal answer. Other processing modules execute other important functions: speech act classification, linguistic information extraction, evaluation of student assertions, speech production with the animated conversational agent, and others which need not be described in this paper.

Previous Empirical Studies of Tutorial Learning

One-to-one tutoring is a powerful method of promoting knowledge construction as has been shown through available empirical studies (Bloom, 1984; Cohen, Kulik, & Kulik, 1982; Corbett, 2001). The vast majority of the tutors in these studies of human tutoring have had moderate domain knowledge and little or no training in pedagogy or tutoring; the tutors were peer tutors, cross-age tutors, or paraprofessionals, but very rarely accomplished tutors. The unaccomplished human tutors enhanced learning with an effect size of .4 standard deviation units (called sigmas), which translates to approximately an improvement of half a letter grade (Cohen et al., 1982). The accomplished human tutors produced effect sizes of 2 sigmas according to Bloom (1984), although the magnitude of this effect should be questioned due the relative small amount of studies that have looked at accomplished tutors.

In the arena of computer tutors, intelligent tutoring systems with sophisticated pedagogical tactics but no natural language dialog produce effect sizes of approximately 1 sigma (Corbett, 2001; VanLehn et al., 2002). Previous versions of AutoTutor have produced gains of .4 to 1.5 sigma (a mean of .7), depending on the learning measure, the comparison condition, the subject matter, and version of AutoTutor (Person et al., 2001; VanLehn & Graesser, 2002). This places previous versions AutoTutor somewhere between an unaccomplished human tutor and an intelligent tutoring system. It might be noted, however, that one recent evaluation of physics tutoring (VanLehn & Graesser, 2002) remarkably reported that the learning gains produced by accomplished human tutors were equivalent to the gains produced in two computer tutors with natural language dialog (Why/AutoTutor and Why/Atlas, a system developed at the University of Pittsburgh). The effectiveness of different tutoring systems clearly requires additional research.

Present Study of Why/AutoTutor

We conducted an experiment that assessed learning gains of Why/AutoTutor, compared with two comparison conditions. Those assigned to the **AutoTutor** Condition learned conceptual physics by participating in a tutorial dialog with Why/AutoTutor for approximately 3-4 hours. Those in the **Read-textbook** condition read textbook chapters on the same Newtonian physics topics covered by Why/AutoTutor, for a comparable amount of study time; the textbook was Hewitt's *Conceptual Physics* (1998). There was also a no-material **Control** condition in which the subjects did not receive any material on conceptual physics. The participants were 35 college students enrolled in a college physics course at Ole Miss, Rhodes College, and the University of Memphis. The participants were randomly assigned to the three conditions, except that twice as many subjects were to be assigned to the AutoTutor condition as in the two comparison conditions. (It should be noted that additional subjects are currently being run, so a larger N will be available soon). Learning gains were assessed by administering a pretest and a posttest that consisted of multiple choice questions. The questions were extracted from or were similar to those in the Force Concept Inventory (Hestenes, Wells, & Swackhamer, 1992). Another method of assessing learning was the quality of their answers to an additional sample of qualitative physics questions, but these data are not reported in the present study.

The experiment included two sessions, approximately 2-3 hours each, one week apart. The first session consisted of a pretest followed by a learning phase, while the second session began with the learning phase and ended with a posttest. Two different test versions (A, B) were counterbalanced across conditions as pre and post tests. Each test has a multiple choice part and a conceptual physics essay part. There were 40 multiple choice items pulled from the Force Concept Inventory (FCI) in each version, A and B. There were 4 conceptual physics questions in each of the two versions of the test.

During the learning phases, participants received either Why/AutoTutor (N=21), Read-textbook (N=8), or Control (N=6). The learning phase of Why/AutoTutor covered 10 conceptual physics questions, such as the example in Figure 1. Each problem took approximately 20 minutes to answer, as the student and AutoTutor collaborative answered the questions. The participants in the Read-textbook condition read the textbook for an approximately equivalent amount of time, as estimated by the tutoring sessions reported in VanLehn and Graesser (2002). VanLehn and Graesser (2002) cover additional details about the tests, learning materials, and methodology.

We computed the proportion of multiple choice questions that were answered correctly on the pretest and posttest. Table 1 presents the means and standard deviations (SD) of the pretests and posttests in the three conditions. The right column in table includes adjusted posttest scores that statistically control for the pretest score; standard errors are in parentheses.

An ANOVA was conducted on the scores, using a 3x2 factorial design, with condition as a between-subject variable and test phase (pre versus post) as a repeated measures variable. There was a statistically significant condition by test phase interaction, $F(2,32) = 10.69$, $p < .01$, $MS_{error} = .005$. The pattern of means clearly showed more learning gains from pretest to posttest in the Why/AutoTutor condition than the other two conditions. An ANCOVA was statistically significant when we analyzed the posttest

scores, using the pretest scores as a covariate, $F(2,31)= 11.37$, $p < .01$. The adjusted posttest scores showed the following ordering among means: Why/AutoTutor > Read-textbook > Control. The effect size (sigma) of the learning gains of Why/AutoTutor was .83 when its pretest served as a control, .97 when the adjusted Control mean served as the control, and 1.02 when the adjusted Read-textbook mean served as the control. These effect sizes are comparable to the intelligent tutoring of systems on physics reported by VanLehn et al. (2002).

Table 1: Proportion Correct on Pretests and Posttests of Three Conditions

Condition	Pretest Mean (SD)	Posttest Mean (SD)	Adjusted Posttest (Std. Error)
AutoTutor	0.602 (.166)	0.739 (.130)	0.738 (.018)
Read-textbook	0.516 (.125)	0.575 (.104)	0.632 (.031)
Control	0.704 (.133)	0.638 (.177)	0.567 (.036)

Two alternative measures of learning gains were computed to show differences between conditions. First, the simple learning gains were computed as Posttest-Pretest. A one-way ANOVA performed on the simple learning gains showed significant differences among conditions, $F(2,32)=10.69$, $p < .01$, $MS_{error} = .009$. As shown in Table 2, and confirmed in follow up planned comparisons, there was the following ordering of means: Why/AutoTutor > Read-textbook > Control. Second, we computed the normalized gain score, a standard that often has been used to report learning gain proportions: [(Posttest-Pretest) / (1-Pretest)]. An ANOVA performed on the normalized gain scores showed the same significant effect, $F(2,32)=18.39$, $p < .01$, $MS_{error} = .042$, and ordering of means.

Table 2: Learning Gains Proportions

Condition	Simple Learning Gains (SD)	Normalized Gain Score (SD)
AutoTutor	0.137 (.108)	0.333 (.231)
Read-textbook	0.059 (.083)	0.105 (.152)
Nothing	-0.067 (.056)	-0.225 (.141)

Conclusions

These results of the present study on qualitative physics follow previous trends in AutoTutor research that have continually shown it to be an effective learning tool (Graesser et al., 2001; Person et al., 2001). Why/AutoTutor consistently outperformed its comparison conditions, in three alternative comparisons that were considered (pretest for Why/AutoTutor, Read-textbook control, and a no learning material Control). These results support the claim that there is something about tutorial dialog in natural language that promotes learning in these constructivist learning environments. We are currently exploring what it is, more precisely, that accounts for the learning gains. Now that we know that learning does occur, we can dissect the potential causes of learning.

Acknowledgments

The Tutoring Research Group (TRG) is an interdisciplinary research team comprised of approximately 35 researchers from psychology, computer science, physics, and education (Visit http://www.autotutor.org). This research conducted by the authors and the TRG was supported by the National Science Foundation (REC 0106965), and the DoD Multidisciplinary University Research Initiative (MURI) administered by ONR 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. Kurt VanLehn, Carolyn Rose, Pam Jordan, and others at the University of Pittsburgh collaborated with us in preparing materials and testing their own intelligent tutoring system with tutorial dialog, called Why/Atlas.

References

Aleven V. & Koedinger, K. R. (2002). An effective metacognitive strategy: Learning by doing and explaining with a computer-based Cognitive Tutor. Cognitive Science, 26.

Bloom, B. S. (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.

Chi, M.T.H., de Leeuw, N., Chiu, M. & LaVancher, C. (1994) Eliciting self-explanation improves understanding. *Cognitive Science, 18*, 439-477.

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

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

Collins, A., Brown, J. S., & Newman, S. E. (1989). Cognitive apprenticeship: Teaching the craft of reading, writing, and mathematics. In L. B. Resnick (Ed.), *Knowing, learning, and instruction: Essays in honor of Robert Glaser* (pp. 453-494). Hillsdale, NJ: Erlbaum.

Corbett, A.T. (2001). Cognitive computer tutors: Solving the two-sigma problem. *User Modeling: Proceedings of the Eighth International Conference* (p. 137-147).

Foltz, P.W., Gilliam, S., & Kendall, S. (2000). Supporting content-based feedback in on-line writing evaluation with LSA. *Interactive Learning Environments, 8*, 111-127.

Forbus, K. (1984). Qualitative process theory. *Artificial Intelligence, 24*, 85-168.

Fox, B. (1993). *The human tutorial dialog project*. Hillsdale, NJ: Erlbaum.

Graesser, A.C., Person, N., Harter, D., & TRG (2001). Teaching tactics and dialog in AutoTutor. *International*

Journal of Artificial Intelligence in Education, 12, 257-279.

Graesser, A. C., Person, N. K., & Magliano, J. P. (1995). Collaborative dialogue patterns in naturalistic one-to-one tutoring. *Applied Cognitive Psychology, 9*, 1-28.

Graesser, A.C., VanLehn, K., Rose, C., Jordan, P., and Harter, D. (2001). Intelligent tutoring systems with conversational dialogue. *AI Magazine, 22*, 39-51.

Graesser, A.C., Wiemer-Hastings, K., Wiemer-Hastings, P., Kreuz, R., & the TRG (1999). Auto Tutor: A simulation of a human tutor. *Journal of Cognitive Systems Research, 1*, 35-51.

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force Concept Inventory. *The Physics Teacher 30*. 141-158.

Hewitt, P.G. (1998). *Conceptual physics* (8th edition). Reading, MA: Addison-Wesley.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge, MA: Cambridge University Press.

Landauer, T.K., Foltz, P.W., Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes, 25*, 259-284.

Moore, J.D. (1995). *Participating in explanatory dialogues*. Cambridge, MA: MIT Press.

Olde, B. A., Franceschetti, D.R., Karnavat, Graesser, A. C. & the Tutoring Research Group (Aug., 2002). The right stuff: Do you need to sanitize your corpus when using latent semantic analysis? *Proceedings of the 24th Annual Conference of the Cognitive Science Society* (pp. 708-713). Mahwah, NJ: Erlbaum.

Person, N. K., Graesser, A. C., Bautista, L., Mathews, E. C., & the Tutoring Research Group (2001). 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* (pp. 286-293). Amsterdam, IOS Press.

Person, N.K., Graesser, A.C., Kreuz, R.J., Pomeroy, V., & TRG (2001). Simulating human tutor dialog moves in AutoTutor. *International Journal of Artificial Intelligence in Education. 12*, 23-39.

Ryder, J.M., Graesser, A.C., McNamara, J., Karnavat, A., & Pop, E. (2002). A dialog based intelligent tutoring system for practicing command reasoning skills. Paper presented at I/ITSEC.

Shah, F., Evens, M., Michael, J., & Rovick, A. (2002). Classifying student initiatives and tutor responses in human keyboard-to-keyboard tutoring sessions. *Discourse Processes, 33*, 23-52.

VanLehn, K. & Graesser, A. C. (2002). Why2 Report: Evaluation of Why/Atlas, Why/AutoTutor, and accomplished human tutors on learning gains for qualitative physics problems and explanations. Unpublished report prepared by the University of Pittsburgh CIRCLE group and the University of Memphis Tutoring Research Group.

VanLehn, K., Jones, R. M. & Chi, M. T. H. (1992). A model of the self- explanation effect. *Journal of the Learning Sciences, 2*(1), pp. 1-60.

VanLehn, K., Lynch, C., Taylor, L.,Weinstein, A., Shelby, R., Schulze, K., Treacy, D., & Wintersgill, M. (2002). In S. A. Cerri, G. Gouarderes, & F. Paraguacu (Eds.), *Intelligent Tutoring Systems 2002* (pp. 367-376). Berlin, Germany: Springer.