Guru: A Computer Tutor that Models Expert Human Tutors

Andrew Olney[1], Sidney D'Mello[2], Natalie Person[3], Whitney Cade[1], Patrick Hays[1], Claire Williams[1], Blair Lehman[1], and Art Graesser[1]

[1] University of Memphis
[aolney|wlcade|dphays|mcwllams|balehman|a-graesser]@memphis.edu
[2] University of Notre Dame
sdmello@nd.edu
[3]Rhodes College
person@rhodes.edu

Abstract. We present *Guru*, an intelligent tutoring system for high school biology that has conversations with students, gestures and points to virtual instructional materials, and presents exercises for extended practice. Guru's instructional strategies are modeled after expert tutors and focus on brief interactive lectures followed by rounds of scaffolding as well as summarizing, concept mapping, and Cloze tasks. This paper describes the Guru session and presents learning outcomes from an in-school study comparing Guru, human tutoring, and classroom instruction. Results indicated significant learning gains for students in the Guru and human tutoring conditions compared to classroom controls.

1 Introduction

Guru is a dialogue-based intelligent tutoring system (ITS) in which an animated tutor agent engages the student in a collaborative conversation that references a multimedia workspace displaying and animating images that are relevant to the conversation. Guru provides short lectures on difficult biology topics, models concepts, and asks probing questions. Guru analyzes typed student responses via natural language understanding techniques and provides formative feedback, tailoring the session to individual students' knowledge levels. At other points in the session, students produce summaries, complete concept maps, and perform Cloze tasks. To our knowledge, Guru is the first ITS that covers an entire high school biology course.

Guru is distinct from most dialogue-based ITSs, such as AutoTutor [1] or Why-Atlas [2], because it is modeled after 50-hours of *expert* human tutor observations that reveal markedly different pedagogical strategies from previously observed novice tutors [3]. Our computational models of expert tutoring are multi-scale, from tutorial modes (e.g. scaffolding), to collaborative patterns of dialogue moves (e.g. information-elicitation), to individual moves (e.g. direct instruction) [4]. However, the importance of tutoring expertise has recently been called into question. In a meta-

adfa, p. 1, 2011.

analysis, VanLehn [5] examined the effectiveness of step-based ITSs and human tutoring compared to no tutoring learning controls matched for content. He reported that the effect sizes of human tutoring are not as large as Bloom's two sigma effect [6]. Instead, the effect sizes for human tutoring are much lower (d = .79), and step-based systems (d = .76) are comparable to human tutoring. Even so, the *relative* influence of expertise on learning outcomes remains unclear and requires more research.

The present study addresses the effectiveness of Guru in promoting learning gains. Specifically, how do learning gains obtained from classroom instruction + Guru compare to classroom + human tutoring and classroom instruction alone? We begin with a sketch of Guru followed by an experiment designed to evaluate the effectiveness of Guru in an authentic learning context, namely an urban high school in the U.S.

2 Brief Description of Guru

Guru covers 120 biology topics aligned with the Tennessee Biology I Curriculum Standards, each taking from 15 to 40 minutes to cover. Topics are organized around *concepts*, e.g. *proteins help cells regulate functions*. Guru attempts to get students to articulate each concept over the course of the session. In this study, a Guru session is ordered in phases: Preview, Lecture, Summary, Concept Maps I, Scaffolding I, Concept Maps II, Scaffolding II, and Cloze Task. Guru begins with a **Preview** making the topic concrete and relevant to the student, e.g. "Proteins do lots of different things in our bodies. In fact, most of your body is made out of proteins!" Guru's **Lectures** have a 3:1 (Tutor:Student) turn ratio [4, 7] in which the tutor asks concept completion questions (e.g., Enzymes are a type of what?), verification questions (e.g., Is connective tissue made up of proteins?), or comprehension gauging questions (e.g., Is this making sense so far?). At the end of the lectures, students generate **Summaries**; summary quality determines the concepts to target in the remainder of the session. For target concepts, students complete skeleton **Concept Maps** which are automatically generated from concept text [8]. In **Scaffolding,** Guru uses a Direct Instruction → Prompt → Feedback → Verification Question → Feedback dialogue cycle to cover target concepts. A **Cloze** task requiring students to fill in an ideal summary ends the session.

Guru's interface (see Figure 1) consists of a multimedia panel, a 3D animated agent, and a response box. The agent speaks, gestures, and points using motion capture and animation. Throughout the dialogue, the tutor gestures and points to images on the multimedia panel most relevant to the discussion, and images are slowly revealed as the dialogue advances. Student typed input is mapped to a speech act category (e.g., Answer, Question, Affirmative, etc.) using regular expressions and a decision tree learned from a labeled tutoring corpus [9,10]. Guru uses speech act category and multiple models of dialogue context to decide what to do next. Thus an affirmative in the context of a verification question is interpreted as an Answer, while an affirmative in the context of a statement like "Are you ready to begin?" is not. Guru uses a general model of dialogue (e.g., feedback, questions, and motivational dialogue) and specific models representing the *mode* of the tutoring session, including

Lecture and *Scaffolding*. The mode models contain specific logic for answer assessment, feedback delivery (positive, neutral, or negative), and student model maintenance consisting of the concepts associated with each topic. A full description of the system is beyond the scope of the current paper.

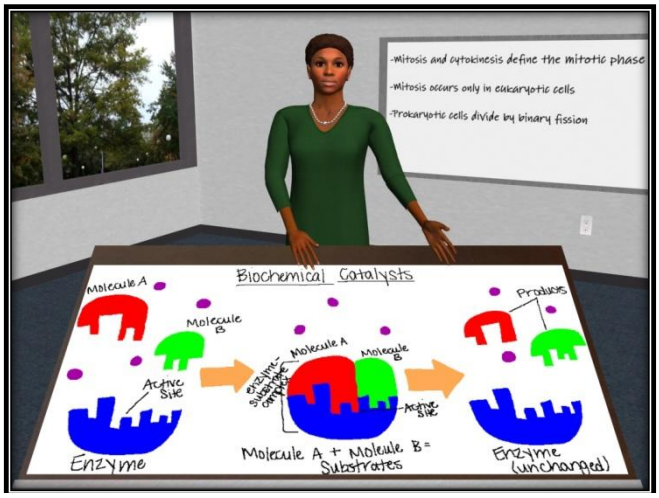

Figure 1. Guru interface

3 Method

Thirty-two tenth graders enrolled in Biology I in an urban U.S. high school participated once a week for three weeks in a three condition repeated-measures study where students interacted with both Guru and a human tutor *in addition* to their regular classroom instruction. Tutored topics were covered in class in the previous week. Space limitations prevent listing the intricate details of the methods. What is important to note is that (1) there were four topics in the study (topics A: Biochemical Catalysts, B: Protein Function, C: Carbohydrate Function, D: Factors Affecting Enzyme Reactions), (2) students received classroom instruction on all four topics, (3) students received additional tutoring for two out of the four topics (A and B), (4) some students were tutored by Guru for topic A and a human tutor for topic B, whereas other students received Guru tutoring for topic B and human tutoring for topic A, (5) tutoring topic (e.g., A or B) was counterbalanced across Guru and the human tutor (6) all students completed pretests, immediate posttests, and delayed posttests on all topics. This design allowed us to (1) compare Guru with human tutoring (e.g., learning gains for topic A vs. B, where topic is counterbalanced across tutors), (2) compare learning gains from tutoring with learning gains from classroom instruction only (gains for A and B vs. C and D), and (3) assess if there are any benefits to classroom instruction alone (i.e., do learning gains for C and D exceed zero).

Knowledge assessments were multiple-choice tests; twelve item pre- and posttests were administered at the beginning and end of each tutoring session to assess prior

knowledge and *immediate learning gains*, respectively. Test items were randomized across pre- and posttests, and the order of presentation for individual questions was randomized across students. Students also completed a 48-item *delayed* posttest the final week. Half of test items were previously used on the immediate pre or posttests, and half were new, with randomized order across students. The researcher who prepared the knowledge tests had access to the topics, the concepts for each topic, the biology textbook, and existing standardized test items. Content from the lectures, scaffolding moves, and other aspects of Guru were *not* made available to the researcher. The researcher was also blind to the tutored condition.

Students and parents provided consent prior to the start of the experiment. Students were tested and tutored in groups of two to four. The procedure for each tutorial session involved (a) students completing the pretest for 10 minutes (b) a tutorial session with either Guru or the human tutor for 35 minutes, and (c) the immediate posttest for 10 minutes. The four human tutors were provided with the topic to be tutored, the list of concepts, and the biology textbook. Each tutor was an undergraduate major or recent graduate in biology. Prior to the study, each tutor participated in a one day training session provided by a nonprofit agency that trains volunteer tutors for local schools. Thus while our tutors might be considered experts in the biology domain, they were not expert tutors.

4 Results

The pretest and immediate and delayed posttests were scored and proportionalized. A repeated measures ANOVA did not yield any significant differences on pretest scores, $F(2, 56) = 1.49$, $p = .233$, so students had comparable knowledge prior to tutoring. Separate proportionalized learning gains for immediate and delayed posttest were computed as follows: (proportion posttest - proportion pretest) / (1 - proportion pretest). This measure tracks the extent to which students acquire knowledge from pre to post. Two scores beyond 3.29 SD from the mean were removed as outliers.

A repeated measure ANOVA on proportional learning gains for the *immediate posttest* was significant, $F(2, 54) = 5.09$, $MSe = .212$, partial eta-square = .159, $p = .009$. Planned comparisons indicated that immediate learning gains for Guru ($M = .385$, $SD = .526$) and human tutoring ($M = .414$, $SD = .483$) did not differ from each other ($p = .846$) and were significantly ($p < .01$) greater than the classroom control ($M = .060$, $SD = .356$). The effect size (Cohen's d) for Guru vs. classroom was 0.72 sigma, while there was a 0.83 sigma effect for the human vs. classroom comparison.

This pattern of results was replicated for the *delayed posttest* (see Figure 2). The ANOVA yielded a significant model, $F(2, 54) = 5.80$, $MSe = .219$, partial eta-square = .177, $p = .005$. Learning gains for Guru ($M = .178$, $SD = .547$) and human tutoring ($M = .203$, $SD = .396$) were equivalent ($p = .860$) and significantly greater ($p < .01$) than the no-tutoring classroom control ($M = -.178$, $SD = .203$). The Guru vs. classroom effect size was 0.75 sigma, the human vs. classroom effect size was 0.97 sigma.

Paired samples t-tests indicated that learning gains on the delayed posttests were significantly lower ($p < .05$) than gains on the immediate posttests for all three condi-

tions, which was expected. There was considerable learning on the delayed posttests for the Guru and human conditions, but not the classroom condition: one-sample *t*-tests indicated that proportional learning gains on the delayed posttests for Guru and human tutoring was significantly greater than 0 (zero is indicative of no learning) but was significantly *less* than zero for the classroom condition.

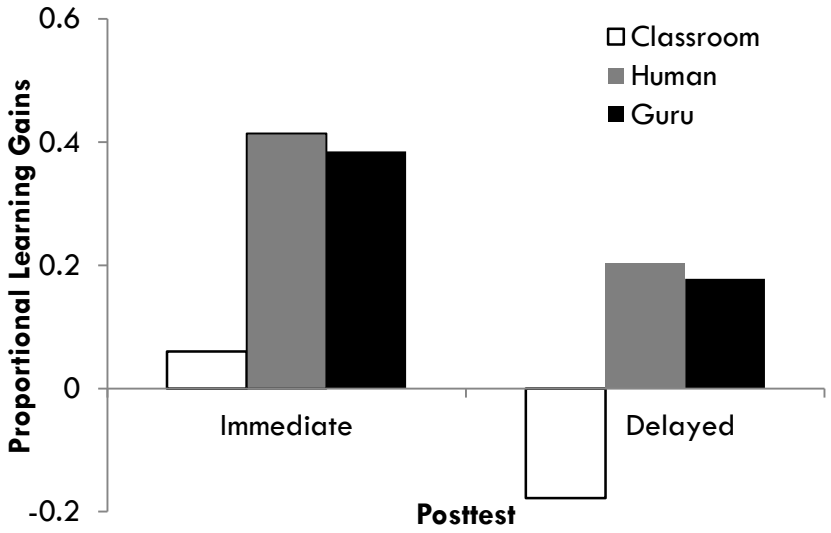

Figure 2. Proportional learning gains

5 General Discussion

These results suggest that Guru is as effective as novice tutors and more effective than classroom instruction only. More importantly, the benefits of tutoring continue after a delay of one to two weeks. Although no differences between Guru and the human tutors were found, there were some limitations to this comparison. First, the human tutors were not able to work one-on-one with 32 students, and so they worked with two to four students simultaneously whereas students worked with Guru individually. However, prior work suggests that the group size may not have detracted from the human tutor condition: Bloom's 2 sigma effect was achieved with groups of 1-3 [6].

Another limitation is that the present human tutors do not meet the same criteria of expertise as the expert tutors on which Guru is modeled, e.g. licensed teachers with considerable tutoring experience (see [11]). Thus the lack of difference between Guru and human tutoring does not clarify Guru's effectiveness vis-à-vis expert human tutors. The .79 effect size for human tutoring reported by VanLehn [5] is highly comparable to the effect size of both Guru and human tutors in the present study, so it is unclear whether an expert tutor under these same conditions would generate significantly greater learning gains. Nonetheless, we are very encouraged by these findings and have preliminary evidence of Guru's efficacy.

Acknowledgment

This research was supported by the National Science Foundation (NSF) (HCC 0834847 and DRL 1108845) and Institute of Education Sciences (IES), U.S. Department of Education (DoE), through Grant 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 NSF, IES, or DoE.

References

1. Graesser, A.C., Lu, S. L., Jackson, G., Mitchell, H., Ventura, M., Olney, A.: AutoTutor: A tutor with dialogue in natural language. Behavioral Research Methods, Instruments, and Computers. 36, 180-193 (2004)
2. VanLehn, K., et al.: The architecture of Why2-Atlas: A coach for qualitative physics essay writing. In: S.A. Cerri, G. Gouarderes, F. Paraguacu (eds.) Proceedings of the Sixth International Conference on Intelligent Tutoring, pp. 158-167. Springer-Verlag, Berlin (2002)
3. Person, N.K., Lehman, B., Ozbun, R.: Pedagogical and Motivational Dialogue Moves Used by Expert Tutors. In: 17th Annual Meeting of the Society for Text and Discourse. Glasgow, Scotland (2007)
4. D'Mello, S.K., Olney, A.M., Person, N.K: Mining collaborative patterns in tutorial dialogues. Journal of Educational Data Mining. 2(1), 1-37 (2010)
5. VanLehn, K.: The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. Educational Psychologist. 46(4), 197-221 (2011)
6. Bloom, B.: The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. Educational Researcher. 13(6), 4-16 (1984)
7. D'Mello, S.K., Hays, P., Williams, C., Cade, W.L., Brown, J., Olney, A.M.: 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. Springer, Berlin / Heidelberg. (2010)
8. Olney, A.M., Cade, W.L., Williams, C.: Generating Concept Map Exercises from Textbooks. In: Proceedings of the Sixth Workshop on Innovative Use of NLP for Building Educational Applications, pp. 111–119. Association for Computational Linguistics, Portland, Oregon (2011)
9. Olney, A.M.: GnuTutor: An Open Source Intelligent Tutoring System Based on AutoTutor. In: Proceeding of 2009 AAAI Fall Symposium on Cognitive and Metacognitive Educational Systems, pp. 70-75. AAAI Press (2009)
10. Rasor, T., Olney, A.M., D'Mello, S.K.: Student Speech Act Classification Using Machine Learning. In: P.M. McCarthy, C. Murray (eds.) Proceedings of 24rd Florida Artificial Intelligence Research Society Conference, p. 275-280. AAAI Press, Menlo Park, CA (2011)
11. Olney, A.M., Graesser, A.C., Person, N.K. Tutorial dialog in natural language. In: R. Nkambou, J. Bourdeau, R. Mizoguchi (eds.) Advances in Intelligent Tutoring Systems, Studies in Computational Intelligence, pp. 181-206. Springer-Verlag, Berlin (2010)