Evaluation of Automatically Generated Hint Feedback

Michael John Eagle
North Carolina State University
Department of Computer Science
890 Oval Drive, Campus Box 8206
Raleigh, NC 27695-8206
mjeagle@ncsu.edu

Tiffany Barnes
North Carolina State University
Department of Computer Science
890 Oval Drive, Campus Box 8206
Raleigh, NC 27695-8206
tiffany.barnes@ncsu.edu

ABSTRACT

This work explores the effects of using automatically generated hints in problem solving tutor environments. Generating hints automatically removes a large amount of development time for new tutors, and it also useful for already existing computer-aided instruction systems that lack intelligent feedback. We focus on a series of problems, after which, previous analysis showed the control group is to be 3.5 times more likely to cease logging onto an online tutor when compared to the group who were given hints. We found a consistent trend in which students without hints spent more time on problems when compared to students that were provided hints.

1. INTRODUCTION

Problem solving is an important skill across many fields, including science, technology, engineering, and math (STEM). Working open-ended problems may encourage learning in higher 'levels' of cognitive domains [1]. Intelligent tutors have been shown to be as effective as human tutors in supporting learning in many domains, in part because of their individualized, immediate feedback, enabled by expert systems that diagnose student's knowledge states [9]. However, it can be difficult to build intelligent support for students in open problem-solving environments. Intelligent tutors require content experts and pedagogical experts to work with tutor developers to identify the skills students are applying and the associated feedback to deliver [6].

Barnes and Stamper built an approach called the Hint Factory to use student data to build a graph of student problem-solving approaches that serves as a domain model for automatic hint generation [7]. Hint factory has been applied across domains [5]. Stamper et al. found that the odds of a student in the control group dropping out of the tutor were 3.5 times more likely when compared to the group provided with automatically generated hints [8]. The hints also affected problem completion rates, with the number of problems completed in L1 being significantly higher for the hint group by half of a standard deviation, when compared to the control group.

This work extends these results by exploring potential causes for these differences. We hypothesized that there would be differences in the amount of time required to solve problems between the students who received hints and the students who did not. We concentrated on the first five problems, before the dropout differences reported in [8]. We found that while there are no differences in total time in tutor, there were some differences in several problems where students in the control group spent significantly more time attempting to solve the problems when compared to the group of students who were provided with hints. This suggests that while both groups spend similar amount of total time in tutor, the group provided with automatically generated hints was able to get further in the tutor. Exploration of the interaction networks [4] for these problems revealed that the control group often spent this extra time pursuing buggy-strategies that did not lead to solutions.

2. THE DEEP THOUGHT TUTOR

We perform our analysis on data from the Deep Thought propositional logic tutor [2]. Each problem provides the student with a set of premises, and a conclusion, and asks students to prove the conclusion by applying logic axioms to the premises. Deep Thought allows students to work both forward and backwards to solve logic problems [3]. Working backwards allows a student to propose ways the conclusion could be reached. For example, given the conclusion B, the student could propose that B was derived using Modus Ponens (MP) on two new, unjustified propositions: $A \rightarrow B, A$. This is like a conditional proof in that, if the student can justify $A \rightarrow B$ and A, then the proof is solved. At any time, the student can work backwards from any unjustified components, or forwards from any derived statements or the premises.

2.1 Data

We perform our experiments on the Spring and Fall 2009 Deep Thought logic tutor dataset as analyzed by Stamper, Eagle, and Barnes in 2011[8]. In this dataset, three different professors taught two semesters each of an introduction to logic course, with each professor teaching one class with hints available and one without hints in the Deep Thought tutor. In the spring semester there were 82 students in the Hint group and 37 students in the Control group; in the fall semester there were 39 students in the Hint group and 83

in the Control group. Students for which application log-data did not exist were dropped from the study; resulting in 68 and 37 students in the Hint group, and 28 and 70 students in the Control group for the first and second semesters respectively. This results in a total of 105 students in the Hint group and 98 students in the Control group. Students from the 6 sections of an introduction to logic course were assigned 13 logic proofs in the deep thought tutor. The problems are organized into three constructs: level one (L1) consisting of the first six problems assigned; level two (L2) consisting of the next five problems assigned; and level three (L3) consisting of the last two problems assigned. We refer to the group that received hints as the Hint group, and the group that did not receive hints as the Control group.

3. RESULTS

In order to investigate the increased rate of drop-out between the hint group and the control group. We concentrate on the first 5 problems from L1 of the Deep Thought Tutor. We focus here as, while the groups started with similar completion and attempt rates, after level one the groups diverge on both completion and problem attempt rates. Since investigation of the interaction networks for these problems revealed that the control group often pursue buggy-strategies, which do not result in solving the problem, we hypothesized that their would be differences in the amount of time spent in tutor between the groups.

We performed analysis on the student-tutor interaction logs. For each student we calculated the summation of their elapsed time per interaction. To control for interactions in which the student may have idled we filtered any interactions that took longer than then minutes. The descriptive statistics for this are located in Table 1, Prob represents the problem number, H and C represent the Hint group and the Control group.

Table 1: Descriptive Statistics for Time (in seconds) Spent in Each Problem

	N		M		SD	
Prob	H	C	H	C	H	C
1.1	104	93	765.89	1245.24	956.41	1614.30
1.2	88	76	761.65	1114.37	911.24	1526.91
1.3	90	67	664.17	1086.09	733.95	2119.19
1.4	87	71	754.60	1266.39	1217.06	1808.53
1.5	84	67	710.62	1423.22	1192.43	2746.54

The large standard deviations are a sign that perhaps this data is not normal. Exploring the data with Q-Q plots reveals that the data is in fact, not normally distributed. This prevents us from performing between-group statistical tests, such as the student's t-test, as our data violates the assumption of normality. To normalize the data, we use a logarithmic transformation (common log) to make the data more symmetric and homoscedastic. Observation of the Q-Q plot and histogram of the transformed data reveal that we had addressed the normality concerns. The results are presented in Table 2.

To test for differences between the two groups on each problem, we subjected the common log transformed data to t-test. The results from this test are presented in Table 3. There are significant differences for problems one, four, and five. The ratio is calculated by taking the difference between the hint group mean and the control group mean. As $\lg(x) - \lg(y) = \lg(\frac{x}{y})$ the confidence interval from the logged data estimates the difference between the population means of log transformed data. Therefore, the anti-logarithms of the confidence interval provide the confidence interval for the ratio. We provide the C:H ratios and confidence intervals in Table 4.

Table 2: Descriptive Statistics After Common Log Transformation

	N		M		SD	
Prob	H	C	H	C	H	C
1.1	104	93	2.63	2.79	0.48	0.55
1.2	88	76	2.59	2.73	0.54	0.54
1.3	90	67	2.62	2.72	0.44	0.48
1.4	87	71	2.66	2.89	0.40	0.41
1.5	84	67	2.55	2.75	0.48	0.60

Table 3: Ratio Between Groups (H:C) in the Original Scale

		95% Confidence Interval			
Prob	Ratio	low	high	p-value	t
1.1	0.69	0.50	0.97	0.03	-2.18
1.2	0.72	0.49	1.06	0.10	-1.68
1.3	0.78	0.56	1.10	0.15	-1.43
1.4	0.58	0.44	0.78	0.00	-3.61
1.5	0.62	0.42	0.93	0.02	-2.31

Table 4: Ratio Between Groups (C:H) in the Original Scale

		95% CI	
Prob	Ratio	low	high
1.1	1.44	1.04	2.01
1.2	1.39	0.94	2.05
1.3	1.27	0.91	1.78
1.4	1.71	1.28	2.30
1.5	1.60	1.07	2.40

In order to explore what these differences mean, we shall transform the data back to our original scale (seconds.) The transformed data is provided in Table 5. These are the Geometric Means, which are often closer to the original median, than they are the mean. The ratios from Tables 3 and 4 are easily interpreted as the log of the ratio of the geometric means. For example in problem 1.4, in the common log scale, the mean difference between hint and control group is -0.23. Therefore, our best estimate of the ratio of the hint time and control time is $10^{-.23} = 0.58$. Our best estimate of the effect of Hint is it takes 0.58 times as many seconds as the control group to complete the problem. The confidence interval reported above is for this difference ratio.

The geometric mean of the amount of seconds needed to solve problem four for the hint group is 0.58 (95% CI: 0.44 to 0.78) times as much as that needed for students in the control group. Stated alternatively, students in the control group spend 1.71 (95% CI: 1.07 to 2.40) times as long as the Hint group in problem four.

Table 5: Geometric Means and Confidence Intervals in Seconds

		95% CI			95% CI	
P	H	low	high	C	low	high
1	428.66	347.14	529.31	618.19	478.60	798.51
2	387.07	297.97	502.82	537.80	405.75	712.82
3	413.80	335.89	509.78	527.18	405.05	686.13
4	454.43	374.38	551.61	778.01	624.48	969.29
5	352.90	278.06	447.89	565.61	405.34	789.24

Exploring the total time spent between all five problems also required a log transformation. The total time spent on the first 5 problems between the hint group ($M = 3.34, SD = 0.4$) and the control group ($M = 3.44, SD = 0.51$) was not significant, $t(198) = 1.41, p = 0.16$. This corresponds to a H:C ratio of 0.81 (95% CI: 0.60 to 1.09), and a C:H ratio of 1.24 (95% CI: 0.92 to 1.66).

4. DISCUSSION

The results of this analysis show that students in the control group are overall not spending significantly more time in the tutor during these first five problems. However, the control does spend significantly more time in some problems compared to the hint group. Problems one, three and four provided students with the automatically generated hints. While problem two and five had no hints for either group. We would expect there to be differences in time to solve for the hint group, and this was the case for problem one. We would also expect that having no hints on problem two would not display an effect, as the second problem is too early to expect differences to emerge between the groups. Problem three is interesting as this problem is the first in which the groups begin to show preferences towards different solution strategies. With the control group preferring to work backwards, and the hint group preferring to work forwards (hints are only available for solutions working forward). Problem four and five, both of which showed significant differences in time spent, showed a large portion of control group student interactions to be perusing buggy-strategies.

This is interesting as the control group is spending at least as much, and often more, time in tutor and yet meeting with less overall success. The control students are not becoming stuck in a single bottleneck location within the problems and then quitting, which would result in lower control group times. The control students are actively trying to solve the problems using strategies that do not work. The hint group is able to avoid these strategies via the use of the hints. The hint group students also develop a preference for solving problems forward, as that is the direction in which they can ask for hints. It is interesting to see that these preferences remain, even when hints are not available.

The effect of the automatically generated hints appear to let the hint group spend around 60% of the time per problem compared to the control group. Or stated differently, the control group requires about 1.5 times as much time per problem when compared to the hint group. These results show that the hints provided by the Hint Factory, which are generated automatically, can provide large differences in how long students need to solve problems.

5. CONCLUSIONS AND FUTURE WORK

This paper has provided evidence that automatically produced hints can have drastic effects on the amount of time that students spend solving problems in a tutor. We found a consistent trend in which students without hints spent more time on problems when compared to students that were provided hints. Exploration of the interaction networks for these problems revealed that the control group often spent this extra time pursuing buggy-strategies that did not lead to solutions. Future work will explore other data available on the interaction level, such as errors, in order to get a better understanding of what the control group is doing with their extra time in tutor. We will also look into the development of further interventions that can help students avoid spending time on strategies that are unlikely to provide solutions.

6. REFERENCES

[1] B. S. Bloom. *Taxonomy of Educational Objectives: The Classification of Educational Goals*. Taxonomy of educational objectives: the classification of educational goals. Longman Group, New York, 1956.

[2] M. J. Croy. Graphic interface design and deductive proof construction. *J. Comput. Math. Sci. Teach.*, 18:371–385, December 1999.

[3] M. J. Croy. Problem solving, working backwards, and graphic proof representation. *Teaching Philosophy*, 23:169–188, 2000.

[4] M. Eagle, M. Johnson, and T. Barnes. Interaction Networks: Generating High Level Hints Based on Network Community Clustering. *educationaldatamining.org*, pages 1–4.

[5] D. Fossati, B. Di Eugenio, S. Ohlsson, C. Brown, L. Chen, and D. Cosejo. I learn from you, you learn from me: How to make ilist learn from students. In *Proceedings of the 2009 conference on Artificial Intelligence in Education: Building Learning Systems that Care: From Knowledge Representation to Affective Modelling*, pages 491–498, Amsterdam, The Netherlands, The Netherlands, 2009. IOS Press.

[6] T. Murray. Authoring intelligent tutoring systems: An analysis of the state of the art. *International Journal of Artificial Intelligence in Education (IJAIED)*, 10:98–129, 1999.

[7] J. Stamper, T. Barnes, L. Lehmann, and M. Croy. A pilot study on logic proof tutoring using hints generated from historical student data. *Proceedings of the 1st International Conference on Educational Data Mining (EDM 2008)*, pages 197–201, 2008.

[8] J. C. Stamper, M. Eagle, T. Barnes, and M. Croy. Experimental evaluation of automatic hint generation for a logic tutor. In *Proceedings of the 15th international conference on Artificial intelligence in education*, AIED'11, pages 345–352, Berlin, Heidelberg, 2011. Springer-Verlag.

[9] K. VanLehn. The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, 46(4):197–221, 2011.