An Evaluation of the Effectiveness of Just-In-Time Hints

Robert G.M. Hausmann, Annalies Vuong, Brendon Towle,
Scott H. Fraundorf, R. Charles Murray, and John Connelly

Carnegie Learning, Inc.
Frick Building, Suite 918
437 Grant Street, Pittsburgh, PA, 15219
{bhausmann,avuong,btowle,sfraundorf,cmurray,
jconnelly}@carnegielearning.com

Abstract. The present study evaluates the effectiveness of *Just-In-Time Hints* (*JITs*) by testing two competing hypotheses about learning from errors. The *tutor-remediation hypothesis* predicts that students learn best when a tutoring system immediately explains why an entry is incorrect. The *self-remediation hypothesis* predicts that learning is maximized when learners attempt to correct their own errors. The *Cognitive Tutor* was used to test these hypotheses because it offers both JITs, which map onto the tutor-remediation hypothesis, and flag feedback, which maps onto the self-remediation hypothesis. To evaluate the effectiveness of JITs, we conducted a naturalistic experiment where learning from older versions of the software, which did not include specific JITs, was contrasted with a later version that included the JITs. The results suggest JITs reduced the frequency of errors; however, this observation was qualified by an aptitude-treatment interaction whereby high- and low-prior knowledge students differentially benefited from JIT availability.

Keywords: Just-in-time help, feedback, naturalistic experimentation, aptitude-treatment interaction.

1 Introduction

One core belief of the intelligent tutoring system (ITS) community is that intelligent tutoring systems are effective because they provide contextually relevant assistance on individual steps, typically in the form of a hint [1]. Either the student can request a hint, or the system can provide assistance based on the student's recent performance. In the case where the system knows about a common student misconception and automatically delivers a hint based on the current (wrong) entry, we refer to that as a *just-in-time* hint, or a *JIT* for short.

A review of the learning literature suggests two hypotheses about maximizing the probability a student will learn from an error. The first, which we call the *tutor-remediation hypothesis,* posits that learning occurs during the explicit remediation of an error. For example, Anderson *et al.* found that students who received explanatory feedback made fewer errors than students who received only error-flagging feedback.

K. Yacef et al. (Eds.): AIED 2013, LNAI 7926, pp. 791–794, 2013.

Students who received explanatory feedback also took less time to complete the task, although the differences did not persist on a long-term assessment [2].

The second hypothesis, which we call the *self-remediation hypothesis*, places more emphasis on students generating their own explanations for mistakes. In general, learning is more effective when students are required to generate or process the to-be-learned material on their own, rather than having it done for them [3].

2 Method

Both the tutor-remediation and self-remediation hypotheses inform the design of intelligent tutoring systems. The *Cognitive Tutor* employs "just-in-time hints," which map onto the tutor-remediation hypothesis. It also offers immediate "flag feedback," which maps onto the self-remediation hypothesis. Although the *Cognitive Tutor* incorporates both features into its design, the relative strength of including a JIT on a specific problem-solving step is an open question. How much learning improvement might we expect from providing a JIT over and above the mere flagging of an incorrect entry? Moreover, does the effectiveness of a JIT depend on student factors, such as the strength of the student's current understanding? The purpose of our study is to contrast the above hypotheses while looking for an aptitude-treatment interaction [4].

To test the learning improvements provided by a JIT, we needed to compare one group of students presented with a JIT in response to a particular set of inputs, to another group of students not presented with a JIT in response to the same inputs. We elected to do this via a natural experiment; in the course of ongoing *Cognitive Tutor* development, we typically add JITs for inputs that are believed to require them. The result is that students using the tutor in year N+1 will see JITs for inputs that did not provide JITs for students in year N. Thus, our method was to identify JITs that:

1. were contained in sections of the tutor that did not change appreciably at the same time the JIT was added; and,
2. had sufficiently detailed student information logged in our database in the year the JIT was added, as well as in the following year.

These conditions led us to study JITs added in 2008 or 2009. From a random sample of one-third of all schools that used *Cognitive Tutor*, we included all 320 students who produced input that did or could have triggered one of the target JITs. We can thus compare students from the year before the JIT was added to students from the following year, with a high degree of confidence that any changes in student performance were due to the addition of the JIT and not to any confounding factors.

3 Results

To better understand the effectiveness of JITs, we analyzed the percentage of errors (i.e., the number of errors / total number of transactions) between the point at which students did or would have triggered a JIT and the point at which they mastered the skill. Students who did not receive a JIT (M = 26%, SD = 16%) made proportionally more errors on subsequent transactions than students who received a JIT (M = 20%; SD = 14%), $t(434) = 3.32$, $p < .01$, d =.40). This result suggests that JITs were helpful in reducing errors, and it also supports the tutor-remediation hypothesis.

To further clarify this result, we conducted an analysis to determine whether high- versus low-knowledge students differentially benefit from the availability of JITs. We explored this using the estimated probability that the student knows a skill, or *p_known*, given past performance. At the moment she received (or would have received) a JIT, the tutor's runtime engine computes this probability, which is represented as the current value of *p_known* in the Bayesian Knowledge Tracing (BKT) algorithm. A student whose probability of mastery was above the median value was considered *High Prior Knowledge* (High PK), whereas a student who was below the median was labeled *Low Prior Knowledge* (Low PK).

We conducted a 2x2 ANOVA to examine an aptitude (High PK vs. Low PK) by JIT availability (No JIT vs. JIT) interaction. There was a significant main effect for prior knowledge, confirming that our two groups were indeed different, $F(1, 432) = 84.18$, $p < .01$. There was no main effect for the version of the system, $F < 1$. More importantly, the main effects were qualified by a marginally significant interaction and a large effect size, $F(1, 432) = 2.90$, $p < .10$, $\eta_p^2 = .17$.[1] High prior knowledge students took slightly longer to master their skills when they did not see a JIT, whereas the reverse was true for the low prior knowledge students (see Figure 1). The low prior knowledge students who saw a JIT required slightly more interactions with the tutor to master their skills than did students who did not see the JIT. This evidence suggests that JITs' effectiveness may depend on students' having sufficient knowledge to comprehend the JIT's intended message.

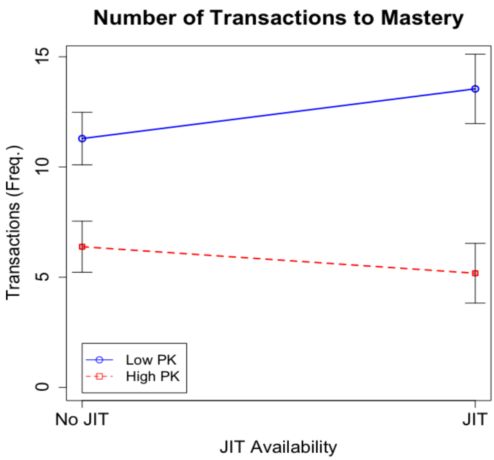

Fig. 1. The interaction between prior knowledge and the exposure to a JIT on the average number of transactions needed to master a skill. Error bars indicate standard error of the mean.

[1] The effect size indicator, partial eta squared (η_p^2), can be interpreted as small when $\eta_p^2 < .06$, medium when $.06 < \eta_p^2 < .14$, and large when $\eta_p^2 > .14$.

4 Discussion

This paper makes two contributions to the ITS literature. First, it demonstrates that the benefits of including a JIT in a sequence of problem-solving steps are contingent on the student's current level of understanding. If students have a sufficiently high understanding of the step, then it is likely that a JIT will help them master the skill more quickly. However, if understanding has not yet reached a certain level, then the JIT is not as effective. The second contribution is methodological. The present analyses represent a "natural experiment" because we were able to manipulate the presence or absence of a JIT depending on the year the software was used. This is analogous to a between-groups experimental design; however, various features of the tutor change between versions because of on-going attempts to improve the software.

To more broadly generalize about the effectiveness of JITs on learning, we would like to extend our sample to include various types of JITs. For example, it would be interesting to extend these analyses by categorizing the types of JITs themselves, such as those that provide simplistic (e.g., "You entered the coordinates with *x* and *y* reversed.") versus conceptually rich (e.g., "Remember that you need to count the area of the base twice, once for the top of the cylinder and once for the bottom.") feedback messages. Different types of JITs may be differentially effective, and certain types may depend more heavily on the strength of the student's current understanding.

In conclusion, we found more evidence in favor of the tutor-remediation hypothesis. Students who were exposed to JITs were more effective in remediating their local errors; however, the long-term impact on learning may be contingent upon the student's current understanding of the skill. Additional work in this area will help build systems that better understand how to use ongoing skill estimates and how to provide students with they help they need.

Acknowledgements. The authors would like to thank the schools that allowed us to collect log files from their students. More importantly, we wish to thank all of the anonymous students who used our software. Special thanks go to Leslie Hausmann for her valuable comments on an earlier draft of this paper.

References

[1] Wood, H., Wood, D.: Help seeking, learning and contingent tutoring. Computers & Education 33, 153–169 (1999)

[2] Anderson, J.R., Corbett, A.T., Koedinger, K.R., Pelletier, R.: Cognitive tutors: Lessons learned. The Journal of the Learning Sciences 4, 167–207 (1995)

[3] Schmidt, R., Bjork, R.: New Conceptualization of Practice: Common Principles in Three Paradigms Suggest New Concepts for Training. Psychological Science 3, 207–217 (1992)

[4] McNamara, D.S., Kintsch, E., Songer, N., Kintsch, W.: Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. Cognition and Instruction 14, 1–43 (1996)