An Ablative Evaluation

Joseph E. Beck\textsuperscript{1, 2}, Ivon Arroyo\textsuperscript{1, 2}, Beverly Park Woolf\textsuperscript{1, 2}, and Carole Beal\textsuperscript{3}
\textsuperscript{1}Computer Science Department, \textsuperscript{2}School of Education, \textsuperscript{3}Psychology Department
University of Massachusetts
Amherst, MA 01003; U.S.A.
Phone: (413) 545-0582; Fax: (413) 545-1249
e-mail: \{beck, ivon, bev\}@cs.umass.edu; cbeal@psych.umass.edu

Abstract

The goal of this project is to use through an intelligent, multimedia computer tutor to increase students' interest in mathematics and their confidence in their ability to learn mathematics. Based on its student model, WhaleWatch selects problems of appropriate difficulty and provides help and instruction as needed. We present the results of a series of two studies. The first study showed that after using the tutor, girls increased in self-confidence at solving mathematics problems. Our second study was ablative, with the tutor's hinting mechanism disabled. We have found that some students respond better (higher motivation and performance) when given less help. Our current goal is to integrate this knowledge into our next design.

1 Introduction and motivation

In the present project, we have developed an intelligent tutor called WhaleWatch to teach fraction concepts. The domain of fractions was selected for three reasons: First, it represents an introduction to more abstract mathematics material and is often regarded as considerably more difficult by students than basic arithmetic. Second, fractions is taught to 11-year-olds in the United States, which is the point at which students, particularly girls, begin to express a dislike of mathematics. Third, there are well-documented and systematic error patterns associated with specific types of fractions problems.

WhaleWatch was specifically designed to support the learning styles that are appealing to girls. Problems involving fractions are presented in the context of the domain of environmental biology, which, of the sciences, is particularly appealing to both girls and boys. The value of learning mathematics is conveyed through the overall goal of assisting the animals, rather than through competition, which girls tend to dislike as the basis for a learning activity.

WhaleWatch also provides individualized instruction and supportive feedback to each student. This individualized instruction is provided by adjusting the level of feedback, controlling the progression through the curriculum, and constructing problems at the “appropriate” level of difficulty. A description of this process can be found in [2]. By making sure the instruction is appropriate for each student, we ensure that students do not become overly discouraged. We hypothesized that this feature should be particularly effective with female students, who are more easily discouraged about their progress in mathematics.
2 Experiments

In Spring 1997, we conducted an evaluation study of WhaleWatch. The participants included 50 students attending two sixth grade classes at an elementary school located in a rural/suburban setting. To assess the students’ changing beliefs about their mathematics ability, a questionnaire was administered in a pre- and post-test design. The questionnaire was drawn from work by Eccles et al. [4]. The questionnaire includes items that tap two dimensions of beliefs about mathematics: the belief that mathematics is useful, and a students self-confidence of her ability in mathematics. Students rate their response to each item on a seven-point scale. The questionnaire has been shown to be highly reliable and to have good psychometric properties [4], and it has been frequently used in prior research on mathematics achievement. This test was given before the students first used the tutor, and after the students completed their final session with WhaleWatch.

Girls who used WhaleWatch increased in self-confidence from 4.9 to 5.3 (significant at P<0.05 via t-test), and boys confidence increased from 5.0 to 5.3. Girls value placed on mathematics increased from 5.4 to 5.9(significant at P≈0.07), while boys’ scores were unchanged. There was no control group in this study, so it was not possible to know what part of the tutor (or if any part) caused this change in scores.

To address these shortcomings we ran an additional study in Spring 1998. The first (experimental) group of students (N=30) used the same version of the tutor as the prior study. The second (control) group of students (N=30) were given a version of WhaleWatch that had its intelligent feedback turned off, so an incorrect response by a student received vague messages such as “Try again”. Other aspects of the tutor were the same. This design is similar in principle to that used by Shute [6] in evaluating StatLady. Table 1 shows the results of this study.

<table>
<thead>
<tr>
<th>Gender</th>
<th>Condition</th>
<th>Before tutor</th>
<th>After tutor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Girls</td>
<td>Experimental</td>
<td>5.40</td>
<td>5.41</td>
</tr>
<tr>
<td>Boys</td>
<td>Experimental</td>
<td>5.29</td>
<td>4.98</td>
</tr>
<tr>
<td>Girls</td>
<td>Control</td>
<td>4.83</td>
<td>4.88</td>
</tr>
<tr>
<td>Boys</td>
<td>Control</td>
<td>5.26</td>
<td>5.62</td>
</tr>
</tbody>
</table>

Table 1. Change in self-confidence from Spring 1998 study.

These data are inconclusive due to a bad split between students in the experimental and control groups. Girls in the experimental group started with a higher confidence than girls in the Spring 1997 study finished with (5.4 vs. 5.3). Since the scale only goes up to 7.0, there is a definite ceiling effect. This makes further increases in confidence very difficult. Girls in the control group started with a similar confidence to those in the Spring 1997 study, but unlike those students (who had the benefits of the feedback), the current group did not exhibit any increase in confidence. This is marginal evidence that feedback is responsible for the effect. Girls mathematics performance using the tutor was significantly better in the experimental than in the control group.

Boys showed a very strange effect, with the control group increasing in confidence (P=0.03, paired samples 2-tailed t-test) and the experimental group decreasing in confidence (P=0.02). Boys in the experimental also showed a significant decrease in the value placed on math, but actual performance with the tutor was comparable across conditions.
3 Conclusions and future work

The results of our ablative study surprised us, and were quite informative. Our hope is that this design will become more common, as there are substantial benefits. By selectively turning off parts of a tutor, it becomes possible to assign credit to various components. Perhaps clever problem generation strategies are actually worthless. More interesting are differential effects. For instance, we have found that some students perform better with little help, while some need all the facilities of an intelligent tutor to guide them. This is somewhat surprising, but the shock is the relative difficulty in partitioning students. We tried to model how well a student would perform in the no-help vs. full-help conditions. Regression models that considered initial confidence, performance on easier problems, and a measure of a student’s level of cognitive development [1] only accounted for 33% of the variance. This was only our first approach, but suggests the problem is non-trivial.

We are now less trusting in results from a single study, particularly with a narrow population of students. Work with the PAT tutor [5] is a good example of a system evaluated with a large, diverse population. Traditionally, evaluations of ITS have been a weak point, but this has been improving. Unfortunately, it can still be difficult to obtain large numbers of subjects for lengthy studies. For example, in our second (1998) study we had four groups of students (girls-experimental, girls-control, boys-experimental, boys-control), but only 46 subjects after attrition. With this number of subjects, it becomes difficult to detect effects. Thus, scaling up the studies we are doing is critical.

In the future, we will better model how much help students require. A better measure of the student's level of cognitive development is a good start in determining how much feedback a student should receive. It is plausible that this is a large factor in how self-directed a student is. We are hesitant to simply ask a student for this information, as our experience with this technique was less-than-successful [3]. Once this is done, we will use machine learning techniques to determine the amount of hinting (if any) a student needs in a given context.

Acknowledgements

This project was supported by National Science Foundation award HRD-9555757 to Beal and Woolf. We thank Mia Stern for her work in designing and developing this tutor, and David Hart and Mary Anne Rea-Ramirez for their help with the evaluation.

References


