The Effects of Error-Flagging in a Tutor on Expression Evaluation

Amruth KUMAR\textsuperscript{a}, Peter RUTIGLIANO\textsuperscript{b}

\textsuperscript{a}Ramapo College of New Jersey, Mahwah, NJ 07430, USA
\textsuperscript{b}New York University, New York, NY

Abstract. We evaluated the effect of providing error-flagging as support for error detection, but not error correction while the student is solving a problem. We found that providing error-flagging in addition to demand feedback during practice learning was no more effective than providing only demand feedback when the tutor did not explicitly mention that errors were being flagged. On the other hand, explaining and providing error-flagging without demand feedback during pre- and post-tests resulted in significantly better scores on pre- and post-tests even though error-flagging did not provide any error-correction support.

Keywords: Demand feedback, Error-flagging, Expression evaluation, Error detection, Error correction

1. Introduction

Demand feedback, error-flagging and immediate feedback are three typical types of feedback provided by tutors. When the different types of feedback provide error detection and error correction, they treat them differently: immediate feedback detects and suggests corrections for errors, whereas demand feedback expects the learner to both detect and correct errors. Error-flagging is a via-media – it detects errors for the learner, but leaves it to the learner whether and how to correct the error.

How does error-flagging stack up against demand feedback? One of the studies with the ACT Programming Tutor [1] found that immediate feedback helped learn the fastest, followed by error-flagging and demand feedback. There was little difference among these groups on tests. Another study with the SQL-Tutor [2] found one of the highest initial learning rates with error-flagging – measured in terms of the probability of violating a constraint after feedback had been provided on the constraint on prior occasions.

We studied error-flagging against demand feedback using a tutor on arithmetic operators. Our study differed from previous studies in the following ways:

- We treated error-flagging as a support mechanism for error-detection only, e.g., in the ACT programming tutor, a student can request advice on an error-flagged step. In our study, error-flagging did not provide any error-correcting advice.
- We provided error-flagging feedback while the student was attempting the problem and demand feedback after the student had submitted the answer – in other words, we compared demand feedback versus error-flagging provided in addition to demand feedback.
- Finally, we considered the effect of error-flagging during testing.

The tutor on arithmetic operators provides demand feedback after the student submits his/her answer. The feedback includes the correct evaluation of the expression. When the tutor is set to provide error-flagging feedback, after the student identifies each step in the evaluation of the expression or enters an intermediate result, the tutor displays them in red if incorrect and green if correct. But, it does not indicate why the answer is incorrect. So, the tutor provides
error-detection but not error-correction support. The student has the option to redo the answer. If the tutor is not set to provide error-flagging, it displays the student’s answer in black. Additional details about the tutor can be found at www.problets.org.

2. Evaluation

In spring and fall 2006, we conducted web-based evaluations of the tutor. We used a between-subjects design, randomly dividing into two groups, the classes that used the tutors. We used the pre-test-practice-post-test protocol for evaluation:

- **Pre-test** consisted of 22 problems, to be solved in 7 minutes. The tutor administered the pre-test, and did not provide any feedback after each answer was submitted. The tutor used the pre-test to initialize the student model.

- **Practice** was adaptive – the tutor presented problems on only those concepts on which the student did not correctly solve problems during pre-test. After the student submitted the answer to each problem, the tutor provided demand feedback including whether the student’s answer was correct, and the correct solution to the problem. Practice session lasted 15 minutes or until the student learned to correctly solve problems on all the concepts, whichever came first.

- **Post-test** consisted of 22 problems similar to those on the pre-test, and presented in the same order. The post-test was limited to 7 minutes. The tutor administered the post-test online, and did not provide any feedback after each answer was submitted.

The three stages: pre-test, practice and post-test were administered back-to-back, with no break in between. The students did not have access to the tutor before the experiment.

In spring 2006, we wanted to evaluate the effect of providing error-flagging in addition to demand feedback during practice. So, the control group received demand feedback, and the test group received error-flagging plus demand feedback during practice. However, when providing error-flagging, the tutor did not explicitly mention that if a step is marked in red, it is incorrect, and if a step is marked in green, it is correct. Neither group received any feedback during the pre-test and post-test.

In fall 2006, we wanted to evaluate the effect of providing error-flagging not only during practice, but also during the pre-test and post-test. So, the control group received demand feedback during practice, and no feedback during pre-test and post-test. The test group received error-flagging plus demand feedback during practice. During pre-test and post-test, it received error-flagging feedback while the student was attempting to solve the expression, but no demand feedback after the student had submitted the answer. The tutor explicitly mentioned that errors were being flagged with color.

We conducted a 2 X 2 mixed factor ANOVA analysis on the problems solved, the total score and the score per problem, with pretest-post-test as the repeated measure, and the type of treatment as between-subjects factor. Our findings for spring 2006 were:

- There was a significant main effect for pre-test versus post-test on the number of problems solved \([F(1,111) = 237.663, p = 0.000]\), the total score \([F(1,111) = 171.162, p = 0.000]\) and the score per problem \([F(1,111) = 40.596, p = 0.000]\).

- There was no significant main effect for the type of treatment and no significant interaction between pre-post and the type of treatment.
Our findings for fall 2006 were:

- There was a significant main effect for pre-test versus post-test on the number of problems solved \( F(1,107) = 102.449, p = 0.000 \), the total score \( F(1,107) = 93.74, p = 0.000 \) and the score per problem \( F(1,107) = 35.603, p = 0.000 \).
- There was a significant main effect for the type of treatment on the total score \( F(1,107) = 13.228, p = 0.000 \), and the score per problem \( F(1,107) = 28.918, p = 0.000 \), but not the number of problems solved \( F(1,107) = 2.037, p = 0.156 \).
- There was no significant interaction between pre-post and the type of treatment.

The average and standard deviation of the total score and score per problem are summarized in Table 2. We found no statistically significant difference in the number of problems solved by the four groups during either the pre-test or practice.

<table>
<thead>
<tr>
<th>Semester</th>
<th>Type of Treatment</th>
<th>Pre-Test Score</th>
<th>Post-Test Score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Total</td>
<td>Per problem</td>
</tr>
<tr>
<td>Spring 2006</td>
<td>Control</td>
<td>5.652</td>
<td>0.549</td>
</tr>
<tr>
<td></td>
<td>Test</td>
<td>4.996</td>
<td>0.542</td>
</tr>
<tr>
<td>Fall 2006</td>
<td>Control</td>
<td>5.441</td>
<td>0.554</td>
</tr>
<tr>
<td></td>
<td>Test</td>
<td>8.074</td>
<td>0.772</td>
</tr>
</tbody>
</table>

Spring 2006 results indicate that providing error-flagging in addition to demand feedback during practice learning is no more effective than providing only demand feedback if the tutor does not explicitly state that errors are being flagged. This could be because: 1) Absent an explicit explanation of error-flagging, the students may not have realized the purpose of color-coded steps on their own. 2) The students were not required to fix errors before proceeding to the next problem. So, students may have simply ignored the feedback.

Fall 2006 results indicate that providing error-flagging during pre- and post-tests results in significantly better scores on pre- and post-tests even though error-flagging did not provide any error-correction support (effect size = 0.65 for the total score and 0.64 for the score per problem.) This was because students with error-flagging support chose to revise their answer far more often (325 revisions on pre-test, 422 on post-test) than those without error-flagging support (70 revisions on pre-test, 71 on post-test). The students with error-flagging did not randomly guess the correct revisions because: 1) they would have spent more time per problem doing so, and would have solved fewer problems, which is not true; 2) the increased score from the pre-test would not have carried over to the post-test, which it has. So, does the increased score signal increased learning? We plan to conduct delayed post-test (without error-flagging) to test this hypothesis.

Acknowledgements: Partial support for this work was provided by NSF under grant CNS-0426021. This work was supported in part by a grant from the Ramapo Foundation.

References