The use of text and process mining techniques to study the impact of feedback on students’ writing processes

Rafael A. Calvo(1), Anindito Aditomo(1,2), Vilaythong Southavilay(1), Kalina Yacef(1),
University of Sydney, Australia,
Email: rafael.calvo@sydney.edu.au, aadi4954@uni.sydney.edu.au, vstoto@gmail.com,
kalina.yacef@sydney.edu.au

(1) The University of Sydney, Australia,
(2) The University of Surabaya, Indonesia

Abstract: Understanding the impact of feedback in complex learning activities, such as writing, is challenging. We contribute a combination of writing environments and data and process mining tools that can provide new ways of measuring this impact. We use the tools in a field experiment in an engineering course (N=45). Responses (timing, amount and types of text changes) were examined using log data and process mining techniques. Two experimental conditions were used: reflective followed by directive feedback (A) and vice-versa (B). We found that both forms of feedback were read multiple times. Students required longer times to respond to reflective, compared to directive, feedback. The type of feedback, however, made little difference to the types of revisions that students performed. Overall, our findings point to the difficulty of encouraging students to reconsider and revise what they have already written.

Introduction

Writing is considered to be a critical form of learning activity at all educational levels. Writing is also a particularly complex activity, and it is generally believed that both novice and experienced writers benefit from feedback provided by others in order to improve their writing. This paper presents a new approach to study the impact of feedback and findings from a field trial in engineering education. This paper examines the impact of different types of writing feedback (directive vs. reflective) on students’ writing process, including the types of revisions students make to their document.

Feedback can be defined as information provided to a person about his/her performance in a task. In educational contexts, the provision of feedback is intended to increase not only a student’s performance, but also the likelihood of learning from the task. Intuitively, feedback should almost always improve learning and performance. Research has shown that the relationship between feedback, performance and learning is nothing but simple (Hattie & Timperley, 2007; Shute, 2008). A meta-analysis found that while feedback improved performance on average, there was a large variation in the effect sizes, and in a third of the studies feedback had a negative impact (Kluger & DeNisi, 1996). To explore possible moderators, Kluger and DeNisi formulated a theory called Feedback Intervention Theory (FIT). FIT draws from control theory and cybernetics to state that feedback interventions cause a person to compare the feedback with a standard or goal. Perceived discrepancies between the feedback and the standard will motivate the person to reduce the discrepancy.

FIT provides a framework to predict the influence of different types of feedback on learning. For instance, feedback with criticism (or praise) towards the learner would divert attention from task relevant processes and can impede learning. Similarly, feedback that highlights one’s performance compared to others (normative feedback, such as grades) would also impede learning. In contrast, feedback that directs attention to the task should facilitate learning, especially if it contains information needed to address the problem highlighted in the feedback (Kluger & DeNisi, 1996, pp. 267-268). Feedback that includes cues about the goal or standard of the task outcome (“goal-setting interventions”) should also increase performance and learning.

Writing, the particular task that we are concerned with here, is much more complex than the typical tasks used in feedback research. Nonetheless, there are some findings consistent with FIT. For instance, FIT would predict that without distinguishing different types of feedback, the effect of writing feedback might be negligible. In line with this prediction, a meta-analysis which lumped together feedback of various kinds found that feedback did not significantly increase the effectiveness of writing interventions on learning (Bangert-Drowns, Hurley, & Wilkinson, 2004, p. 47). But when types of feedback are distinguished, FIT would predict that their impact would differ. Not many studies have examined this issue, but there is some supporting empirical evidence. One study by Nelson and Schunn (2009) examined correlations between features of peer feedback and the likelihood of the feedback being implemented. The writing task was an essay in an undergraduate, introductory course on history. This study found that task-focused feedback (such as those that included specific solutions or specific location of problems) predicted implementation, whereas feedback that focused on the writer (those with affective language such as praise and criticism) did not.
To conjecture further about the role of feedback in writing, we draw from a cognitive model of writing proposed by Bereiter and Scardamalia (1987). They proposed that writing could occur in two different modes: knowledge telling and knowledge transforming. In knowledge telling, the composition process begins with the writer picking up topic and genre cues from the task description, and writing down knowledge from memory activated by these cues. Text already produced becomes an additional source of cues to retrieve knowledge from memory. This process of memory retrieval cued by the task description and text already produced is repeated until the writer feels he/she no longer has any relevant knowledge (or until time or space constraints are met).

The knowledge transforming mode of writing is more complex. It involves the construction and continual reconstruction of a content problem space (what to say) and a rhetorical problem space (how to say it) (Bereiter & Scardamalia, 1987). The two problem spaces interact, with output from one feeding into the other. For instance, a writer in this mode would think about whether the produced text, in its current form (a rhetorical issue), conveys what they intend to say (a content issue). The writer would also think about whether others, or she herself, believe what the text is saying. This may change the way she thinks about the topic, which in turn may prompt her to find a different way to express her new view.

How does writing feedback come into play? Feedback that points to the writing task (as opposed to the writer) has the potential to prompt a writer to reconsider what she/he has written and/or how it was written. In other words, feedback could prompt processes associated with the knowledge-transforming mode of writing, as reflected in more revisions that go beyond cosmetic text changes and also in the greater time lag between feedback and first revision. However, we conjecture that certain types of (task focused) feedback may be more effective than other types at prompting knowledge transforming processes. For instance, feedback that contains specific instructions (what we call here directive feedback) may prompt a writer simply to correct the specific problems, without much consideration. In contrast, feedback that asks students to connect problems in their text with broader content or stylistic issues (what we call here reflective feedback) may prompt more substantial revisions. In this study, we examine two approaches for giving task-focused feedback: directive and reflective. Directive feedback tells the student that there is a problem in the text that needs addressing. Reflective feedback asks the student to consider whether there is a problem.

Method

Research questions and hypotheses
The main question addressed in this study is: “What are the different impacts of reflective and directive feedback on students’ writing process in terms of the timing, amount and types of text changes performed?” As mentioned above, directive feedback explicitly informs a student that their text contains a specific problem, and instructs them to address the problem. Reflective feedback merely suggests the possibility of a problem, but asks the student to decide whether there is a problem that needs to be addressed (see examples in Table 2).

We conjecture that reflective feedback is more effective than directive feedback in prompting students to reconsider their ideas and revise what they have written (processes consistent with a knowledge-transforming mode of writing). More specifically, we predicted that reflective feedback would prompt more deletion of words. Reflective feedback would not necessarily prompt more addition of words, because adding words (expanding a text) can be performed without much reconsideration of text already produced.

Study design and participants
The participants were (N=45) undergraduate and postgraduate engineering students from The University of Sydney. All participants were enrolled in a project-based course where the main activity was to develop a web application. Students had to individually write a proposal that would become the basis of their software development project. The semester was 13 weeks long; the writing activity occurred in the first half of the semester and spanned 32 days.

Table 1. Timeline of the writing assignment

<table>
<thead>
<tr>
<th></th>
<th>Day 0</th>
<th>Day 10</th>
<th>Day 12</th>
<th>Day 17</th>
<th>Day 19</th>
<th>Day 27</th>
<th>Day 32</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group A (n=22)</td>
<td>Assignment opened</td>
<td>1st draft submitted</td>
<td>Directive feedback sent</td>
<td>2nd draft submitted</td>
<td>Reflective feedback sent</td>
<td>Peer feedback sent</td>
<td>Final deadline</td>
</tr>
<tr>
<td>Group B (n=23)</td>
<td>Reflective feedback sent</td>
<td>Directive feedback sent</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In this context, we conducted a crossover field experiment where students were randomly assigned to receive either reflective (n=22) or directive (n=23) tutor feedback in response to the first drafts of their assignment. The first draft was worth 3% of the course mark. Students then had five days to revise before submitting their second draft, which was not assessed. Following this, students who initially received reflective feedback were provided with the directive feedback for their second draft (and vice versa). The crossover setup
helped satisfying ethical requirements by ensuring that students were not systematically disadvantaged by the type of tutor feedback. In addition, after the second tutor feedback, the students had to review their peers’ second drafts. The final submission was worth 5% of the course mark.

**Tools and measurement**

**The writing environment**
Writing activities were managed through iWrite (Calvo, O’Rourke, Jones, Yacef, & Reimann, 2011). Students write on Google Docs, a cloud-based application, and revisions of the documents are retrieved using Google’s API. Tools and heuristics developed to automatically recognize collaborative writing activities (Southavilay, Yacef, & Rafael A Calvo, 2010) was used here to explore what students did after reading the tutor’s feedback.

**Tutor feedback**
The tutor feedback was prepared by the three instructors of the course (who each generated feedback for roughly 1/3 of the students). Feedback was written within the reviewers’ page in iWrite. In order to better align the feedback provided by the three instructors, feedback samples were kept in a shared document. An email announcing that feedback was available at a certain webpage was sent to all students simultaneously. Students were then able to access the feedback on the same interface. Between 3 and 6 feedback items were provided to each student in each phase (M=4.4, SD=1.12). Reflective feedback items had M=96.52 (SD=26.10) number of words at the first release and M=86.86 (SD=23.44) number of words at the second release. Directive feedback items had M=63.77 (SD=16.78) at the first release and M=58.47 (SD=20.66) at the second one.

<table>
<thead>
<tr>
<th>Reflective feedback</th>
<th>Directive feedback</th>
</tr>
</thead>
<tbody>
<tr>
<td>Please think whether including a list of users’ needs, features and design would enhance the clarity of your individual module.</td>
<td>Write a new description of the users’ needs, they are not clear.</td>
</tr>
<tr>
<td>Sometimes spelling mistakes can distract readers from the gist of your argument or message. You may want to check any spelling errors in your document.</td>
<td>Fix the following Spelling problems: [list of problems]</td>
</tr>
</tbody>
</table>

**Data analysis**

**Log preparation**
While users write, Google Docs (GD) saves the document frequently and stores all its revisions. This feature provides a history of the document, where each version of the document can be retrieved with its timestamp. All major revisions available for the 45 documents (an average of N=7.76, SD=4.33 per document) were used. The content of the revisions was used for detecting the type of text change operations. In process mining terms, the history each document is a process instance (case),

In addition, based on the log file of iWrite we used a log file containing all the students’ interactions with the website, with a timestamp each time they accessed the feedback page. This log revealed when and how often individual students access and read their feedback. The feedback-reading log also has one process for each document (student), in total of 45 processes. The revision and feedback-reading logs were merged for each student. A process mining tool, ProM (ProM., 2010) was used to analyze how students accessed feedback and revised their documents. A Dot Chart Analysis (Song & van der Aalst, 2007) was implemented as a plug-in in ProM to extract a snapshot of student activities of revising their documents and accessing the feedback. The result is shown in Figure 1.

**Detecting text change operations**
The model developed by Boiarsky (Boiarsky, 1984) was used to analyze semantic changes in the writing process. In particular, the 12 types of text change operations (Southavilay, Yacef, & Rafael A Calvo, 2010) shown in Table 3 were used. Text change operations at both paragraph and document levels were considered. At the paragraph level, beyond surface changes, the other types of text changes are listed in Table 3.

A text comparison utility was used to identify text change operations (Southavilay, Yacef, & Rafael A Calvo, 2010). The utility is based on the Unix Diff utility, which takes two revisions of text and produces a difference statement with insertions, deletion, and replacements between the two. The text comparison utility uses both paragraph and word-differencing algorithms to detect text changes operations at paragraph and word levels. For each document each two consecutive revisions were compared and the differences computed. First,
the utility uses the paragraph differencing algorithms to discover the addition of new paragraphs, the deletion of existing paragraphs, and the alteration of existing paragraphs. Based on the paragraph-differencing algorithm, we detect text change operations of adding, deleting, moving/reordering, merging, and splitting of existing paragraphs. For each altered paragraph in the later revision, the utility then uses the word differencing algorithm to compare it to the corresponding paragraph in the former revision in order to detect text change operations of moving/reordering, replacing, inserting, deleting, and appending words in the altered paragraphs.

Table 3. Types of text changes automatically identified

<table>
<thead>
<tr>
<th>Text structure</th>
<th>Content change within individual paragraphs (i.e. word level changes)</th>
<th>Content change at the paragraph level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Moving/reordering paragraphs</td>
<td>• Moving or reordering words in a paragraph</td>
<td>• Adding new paragraphs</td>
</tr>
<tr>
<td>Merging paragraphs</td>
<td>• Replacing words in a paragraph</td>
<td>• Deleting paragraphs</td>
</tr>
<tr>
<td>Splitting paragraphs</td>
<td>• Inserting words in the middle of paragraphs</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Deleting words in a paragraph</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Adding words at the end of a paragraph</td>
<td></td>
</tr>
</tbody>
</table>

Findings and discussion

The dotted chart (Figure 1) shows the timing of each student’s access to the tutor feedback, and also the timing of revisions performed in Google Docs. Each row represents a process of one student revising his or her document and accessing the tutor feedback. One “column” (in between white vertical lines) represents a period of one week. Directive feedback is represented by triangles, reflective feedback by squares, and revisions by dots. Group A (who received directive feedback first) occupies the upper half of the chart.

We refer to the period between first and second tutor feedback as Phase 1, and the period between the second feedback and final submission as Phase 2. The black line represents a date when the first tutor feedback sent and the brown line represents a date when the second feedback sent. Several observations can be made from Figure 1. While the feedbacks were sent at the same time, students accessed them at different times (most accessed the feedback on the same day they received it, but some took one or more days). Also, almost all students accessed the feedback multiple times (except for one student who accessed the feedback only once). On average students in Group A accessed the directive feedback 5.8 times and the reflective feedback 4.6 times. In Group B, on average students accessed the reflective feedback 6.3 times, and the directive feedback 4.9 times. Students in both groups revisited the first tutor feedback during Phase 2 (i.e. after receiving their second feedback), as can be seen from the mix of red triangle and blue squares on the dot chart. Hence, students’ behaviors in Phase 2 need to be seen as potentially influenced by both types of tutor feedback. An examination of the timing the changes shows that students took longer to respond to reflective feedback than to directive feedback (1 day longer in Phase 1, and 2 days longer in Phase 2). While not predicted theoretically, this difference is understandable, as reflective feedback requires students to think for themselves, as opposed to being a direct instruction.

In order to see how students revised their documents according to feedback types, we analyzed four types of revision types (adding paragraphs, deleting paragraphs, adding words and deleting words in existing paragraphs) and the corresponding number of words added and deleted. For each phase, we compared the
number of words added and deleted in the four types of revision types in the two experimental conditions. In order to distinguish between headings (not considered in the analysis) and content paragraphs, we defined a paragraph as containing at least 7 words. This resulted in several findings.

First, not many students performed major revisions, especially in Phase 1. For instance, in Phase 1, only 13 students from Group A and 12 from Group B added new paragraphs, and only 10 from Group A and 9 from Group B added words to existing paragraphs. Even fewer students revised their documents by deleting words: in Phase 1, only 4 students from Group A and 5 from Group B deleted paragraphs, and only 8 from Group A and 3 from Group B deleted words in existing paragraphs. In other words, most of the students in both groups performed only minor revisions in Phase 1. In Phase 2, more students did revise their documents. They mainly did so by adding paragraphs and by adding words to existing paragraphs, rather than deleting paragraphs or deleting words within paragraphs. Over the two phases, only about half of the students (in both groups) performed revisions by deleting words. This indicates that only about half of the students reconsidered and revised what they have already written (which are indicators of knowledge transforming writing processes).

A second set of observations are related to the differences, or lack thereof, between Groups A and B. Although in terms of average number words (see Table 4) Group B seemed to have performed more additions and deletions (in both phases), the box plots above show that there is more variation within rather than between the groups. That is, the difference in the average seemed to be inflated by several students who performed much more extensive revisions than most other students. This difference of amount of revision between the groups largely disappeared if we consider the median (as shown in the box plots). Groups A and B are also similar in terms of the number of students performing the three of the four types of revisions shown in the box plots. There seems to be a difference between the groups in Phase 1, in terms of the deletion of words in existing paragraphs: 8 students in Group A, compared to 3 in Group B (although these 3 students made more extensive deletions compared to the 8 students from Group A).

Together, these observations suggest that the feedback (both reflective and directive) failed to prompt major revisions for most of the students. This is an important pedagogical point: that students have the liberty to take into account, or not, their tutors' feedback. Most students did not revise extensively, despite the feedback given, the time students had to address the feedback, and also the fact that students accessed the feedback several times. Furthermore, the observations indicate that our theoretical predictions about the impact of different feedback types were not supported by the data. This could have been due to several reasons which are more to do with the methods, rather than the hypotheses themselves. One possible explanation is that not all of the reflective feedback was of the same quality: an initial inspection of the reflective feedback items indicated that some items were more directive (instructing students to do specific things). Another possible reason for the lack of difference between the groups' revision behavior is that too few students engaged in extensive revisions during Phase 1, mainly because the students had little incentive to perform those revisions (as indicated above, students' behavior in Phase 2 cannot be taken as indicators of the influence of different feedback types).

References

Acknowledgments
This project has been funded by Australian Research Council DP0986873 and a Google Research Award.