Analysis of a Large-scale Formative Writing Assessment System with Automated Feedback

Peter W. Foltz
Pearson and Univ. of Colorado
4940 Pearl East Circle,
Suite 200
Boulder, CO, 80301 USA
Peter.foltz@pearson.com

Mark Rosenstein
Pearson
4940 Pearl East Circle,
Suite 200
Boulder, CO, 80301 USA
Mark.rosenstein@pearson.com

L@S 2015, March 14–18, 2015, Vancouver, BC, Canada.
ACM 978-1-4503-3411-2/15/03.
http://dx.doi.org/10.1145/2724660.2728688.

Abstract
Formative writing systems with automated scoring provide opportunities for students to write, receive feedback, and then revise essays in a timely iterative cycle. This paper describes ongoing investigations of a formative writing tool through mining student data in order to understand how the system performs and to measure improvement in student writing. The sampled data included over 1.3M student essays written in response to approximately 200 pre-defined prompts as well as a log of all student actions and computer generated feedback. Analyses both measured and modeled changes in student performance over revisions, the effects of system responses and the amount of time students spent working on assignments. Implications are discussed for employing large-scale data analytics to improve educational outcomes, to understand the role of feedback in writing, to drive improvements in formative technology and to aid in designing better kinds of feedback and scaffolding to support students in the writing process.

Author Keywords
Data mining; student actions log analysis; automated writing evaluation; computer assisted instruction; machine learning; mixed effects models.

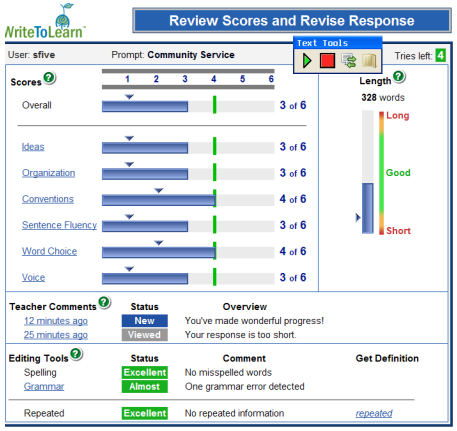

Figure 1. Student feedback screen from WritetoLearn

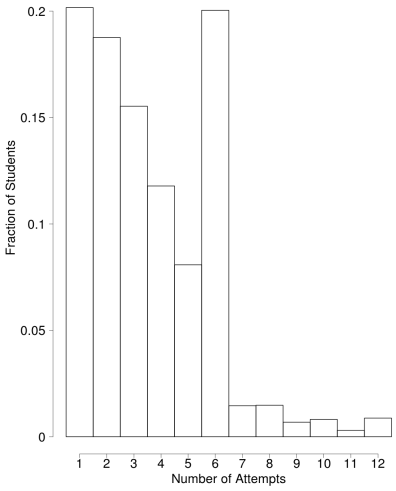

Figure 2. Number of attempts (revisions made by students)

Introduction

Automated scoring of writing, or Automated Essay Scoring (AES), provides the ability to analyze student writing and provide scores and feedback instantly. Studies of AES systems have shown that the scoring of such systems can be as accurate as human scorers (e.g., [1, 2, 3, 4]), can score on multiple traits of writing (e.g., [4]), and can provide feedback on content [3]. While much of the focus in the evaluation of AES has examined the accuracy of the scoring and the different types of essays that can be scored, AES has wide applicability to formative writing, where evaluation can focus on how it aids student learning. By providing instantaneous feedback to students, AES supports the teaching of writing strategies based on detecting the types of difficulties students encounter.

In a formative writing assessment system, all student writing is performed electronically, automatically scored and recorded. Thus, there is a record of all the student actions and all feedback they received. This archive permits continuous monitoring of performance changes in individuals as well as across larger groups of students, such as classes or schools. Teachers can analyze the progress of each student in a class and intervene when needed. In addition it now becomes possible to chart progress across the class in order to measure curricula and teaching effectiveness as reflected in student writing performance scores.

Automated formative assessment of writing provides a rich data set to examine the changes in writing performance and system features that influence that performance. The goal of the present work is to analyze how changes in student writing performance are influenced by features of the system and identify features that promote improvement. We describe instances of applying data mining to components of the formative writing process to investigate specific classes of questions about how a formative system is currently being used, its efficacy, and how understanding current use yields suggestions on ways to improve learning, both through improving the system implementation and by introducing direct interventions aimed at the student during their use of the system.

Method

The formative writing assessment system used for the analyses was WriteToLearn™, a web-based writing environment that provides students with exercises to write responses to narrative, expository, descriptive, and persuasive prompts (see Figure 1). Students use the software as an iterative writing tool in which they write, receive feedback and then revise and resubmit their improved essays. The automated feedback provides an overall score and individual trait scores for aspects of the writing such as "ideas, organization, conventions, word choice, and sentence fluency". Evaluations of WriteToLearn have shown significantly better reading comprehension and writing skill resulting from two weeks of use as well as validating the system scores being as reliable as human raters.

Data

The data comprised two large samples of student interactions with WriteToLearn collected from U.S. adoptions of the software. One set comprised approximately 1.3 million essays from 360,000 assignments written by 94,000 students collected over a 4 year period. The second set represented approximately 62,000 student sessions with nearly 900,000 actions. The data consisted of student essays

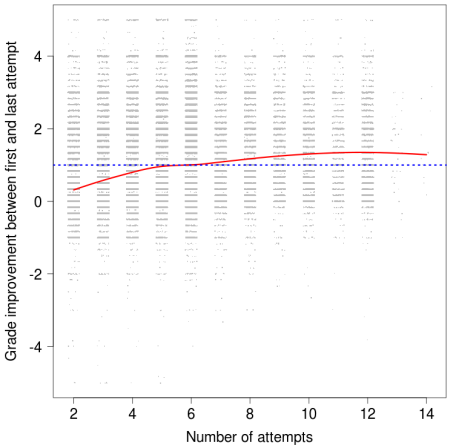

Figure 3. Change in student grade based on number of revisions

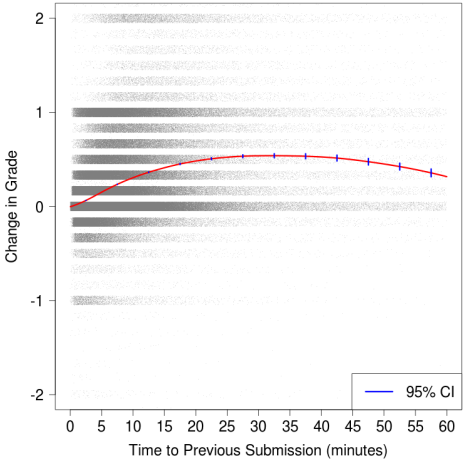

Figure 4. Change in student grade based on time spent between revisions

and a time-stamped log of all student actions, revisions and feedback given by the system. Essays were recorded each time a student submitted or saved an essay, resulting in a record of each draft submitted. The essays were written to approximately 200 pre-defined prompts. No human scoring was performed on these essays. All essay scores were generated by automated scoring, with the prediction performance of the models validated against human agreement from test sets or using a jack-knife procedure.

Results

Number of revisions made by students per prompt

WriteToLearn is designed to support a rapid cycle of write, submit, receive feedback and revise. This cycle is one of the key differentiators of automated formative writing from standard classroom writing practice, where human scoring of essays is time consuming so students can not receive immediate feedback. Thus, it is critical to determine how often students submit and revise essays and determine the factors and time paths that lead to greatest success. Figure 2 shows the distribution of submissions made by students per writing prompt. The distribution shows that nearly equal proportions of students submit a single attempt as submit the default six submissions (approximately 20%). These results indicate that students are taking advantage of revising essays and resubmitting for feedback.

While we see evidence that students are revising their essays, it is important to know whether the revisions result in improved essay quality. Thus, we examined the impact of number of attempts on student performance. In Figure 3 the gray dots show the difference in score between the last and first attempt for an individual student, with essay scores on a six point scale. The red line shows a locally weighted regression curve of the data, indicating about a one score point improvement with 5 or 6 attempts.

Time spent between revisions

We can further investigate the impact on student performance of the time spent writing before requesting feedback (the best allocation of time among the write, submit, feedback, and revise phases). The change in grade shown in Figure 4 indicates that the improvement in writing score generally increases up to about 25 minutes at which point it levels off and begins to drop. We further see that most of the negative change (essays receiving a lower score than the previous version) occurs with revisions of less than five minutes. The results suggest that there is an optimal range of time to spend revising before requesting additional feedback.

Session Analysis – Student Actions

The results above show that students do revise their essays and the time pattern of essay revision suggests that the greatest improvement in writing follows when one revision to the next occurs between 15 to 25 minutes. What the results don't indicate is which actions the students are taking between submissions, and what patterns of actions best improve their writing in those intervals. Thus, we examine performance at the level of actions taken by students within a session.

Actions in the session can include such tasks as logging in/out, submitting an essay for assessment, performing spell or grammar checks, requesting additional formative information about writing (such as examining scoring rubrics, investigating how to improve writing on

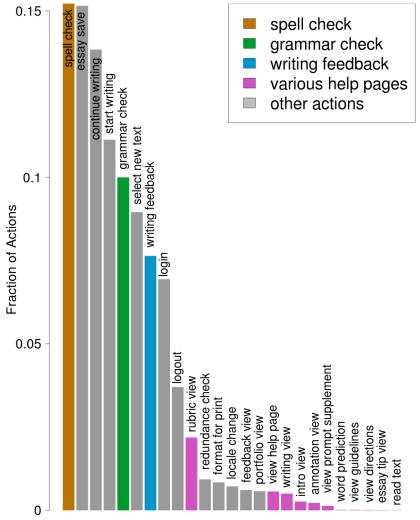

Figure 5. Proportion of student actions taken during writing sessions

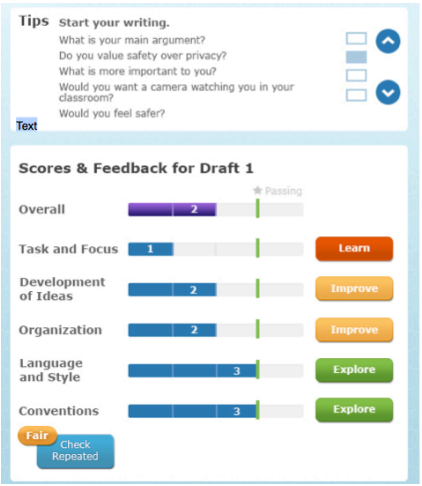

Figure 6. Revised user interface to provide more student writing guidance

different traits, etc.). While there were a median of 15 actions per session, the relative frequency of types of actions performed allows us some insight into the use of features and are shown in Figure 5.

Approximately 15% of the actions were spell check and another 15% were saving the essay (as backup). What is quite noticeable is that almost all the infrequently used actions are various types of request for guidance, such as providing information on the scoring rubric, and suggestions on ways to improve their writing on different writing traits. This suggests that these features needed to be better integrated and made more readily available or automatically provided to students when their need is detected.

Revision of the user interface

Based on the log file analysis, usability results, as well as feedback from customers, a revised user interface was developed which provides additional student feedback and guidance (see Figure 6) Ongoing research is now analyzing how the new interface changes student behavior and performance compared to the original interface.

Conclusions

Large-scale implementations of formative writing provide rich sets of data for analysis of performance and effects of feedback. Automated scoring of writing allows monitoring of student learning as students write and revise essays within these implementations. By examining the log of student actions, the amount of time taken, and the changes in the essays, one can monitor the quality of the students' writing. In assessing writing, the focus has often been put on the product (e.g., the final essay). The analysis of over 1.3M student draft submissions, makes it possible to track the process the learners take to create the product. This analysis allows interventions to be performed on the process of writing rather than just the product. This study further illustrates how data mining can provide new ways of thinking about collecting evidence of system and student performance and uncover patterns that may not be apparent from watching individual students or classrooms.

The overall findings validate a key tenet of formative writing; students improve with revisions and based on feedback from the system. The approach allows examining the changes in learning and the effects of the feedback on writing performance. The data further permits us to discover, prioritize and address concerns as they arise and determine which changes are most likely to improve the students' experience and their ability to sharpen their writing skills.

References

[1] Burstein, J., Chodorow, M., & Leacock, C. (2004). Automated essay evaluation: The Criterion Online writing service. *AI Magazine, 25*(3), 27-36.

[2] Foltz, P. W., Streeter, L. A., Lochbaum, K. E., & Landauer, T. K (2013). Implementation and applications of the Intelligent Essay Assessor. *Handbook of Automated Essay Evaluation*, M. Shermis & J. Burstein, (Eds.). Pp. 68-88. Routledge, NY. NY

[3] Landauer, T. K, Laham, D. & Foltz, P. W. (2001). Automated essay scoring. *IEEE Intelligent Systems.* September/October

[4] Shermis, M. and Hamner. B. (2012). Contrasting state-of-the-art automated scoring of essays: Analysis. In *Paper presented at Annual Meeting of the National Council on Measurement in Education*, Vancouver, Canada, April.