Tracking Student Learning in a State-Wide Implementation of Automated Writing Scoring

Peter W. Foltz
Pearson & University of Colorado
4940 Pearl East Circle, Suite 200
Boulder, CO, 80301
Peter.Foltz@pearson.com

Mark Rosenstein
Pearson
4940 Pearl East Circle
Boulder, CO, 80301
Mark.Rosenstein@pearson.com

Abstract

Student essays provide rich information about the students' knowledge gain and the writing skills acquired. However, hand-scoring is time consuming and does not lend well to large-scale data analyses. Automated scoring of the writing allows monitoring and feedback for individual students as well as tracking changes in performance at district and state levels. We describe an operational implementation of a state-wide formative writing assessment and methods to analyze essays to track student gains as a result of feedback from the assessment. Such approaches become more critical with greater availability of MOOCs and other online learning systems that incorporate online writing.

1 Introduction

The introduction of the Common Core Standards in U.S. K-12 education has led to greater emphasis on integrating reading comprehension and writing skills through student activities such as summarization, writing in response to texts, and demonstrating critical thinking in essays. Although most essays are now being composed with word processors, they are often still submitted in paper form. The student writing is then typically assessed by hand by individual teachers with scores recorded in grade-books and comments interspersed in the essays which are finally returned to students. Thus, while there may be ways to observe changes in a student's attainment by examining performance on each essay, there is little or no trail to track student learning at individual or at the larger group levels.

While large-scale educational data analytics have focused on data from tutoring systems, log-files, gaming, and student information systems, the increased use of writing presents an opportunity to apply large-scale analytics to writing. By providing a unified database for collecting student essays and by employing techniques to automatically score the essays, students can be provided with immediate feedback on their learning. In addition, teachers, administrators and policy makers can track learning at the student, class, school, district and state level. From a data analysis perspective, formative assessment of writing provides a rich data set to examine the changes over time and practice in writing performance and to help in understanding the features of the instructional system that promote improved performance. This paper describes the application of automated writing assessment in a statewide implementation of formative writing, portions of which were presented previously [1]. It then describes approaches used to analyze changes in student writing during the first year of operation.

1.1 Automated assessment of writing

Learning to read and write well results from doing a lot of reading and writing. However, while time on task is a very strong predictor of performance gains in reading and writing, receiving timely feedback is critical [2,3]. Writing provides a rich source of information about student abilities, including comprehension skills, content knowledge, and language ability, and applying automated analyses permits quantifying these skills. Automated assessment of writing has become more readily accepted with multiple systems available for scoring writing both for summative as well as for formative assessment (see [4] for a review). As a formative tool, automated writing assessment can provide students with more opportunities to practice reading comprehension and writing. The technology also allows accurate, immediate individualized feedback that can address the content as well as the surface-level features of their writing [5]. This allows a degree of personalization that embeds practice and assessment within natural performance tasks. Additionally, because all writing is being performed electronically and is scored and recorded automatically, it permits monitoring of performance changes in individual as well as large groups of students.

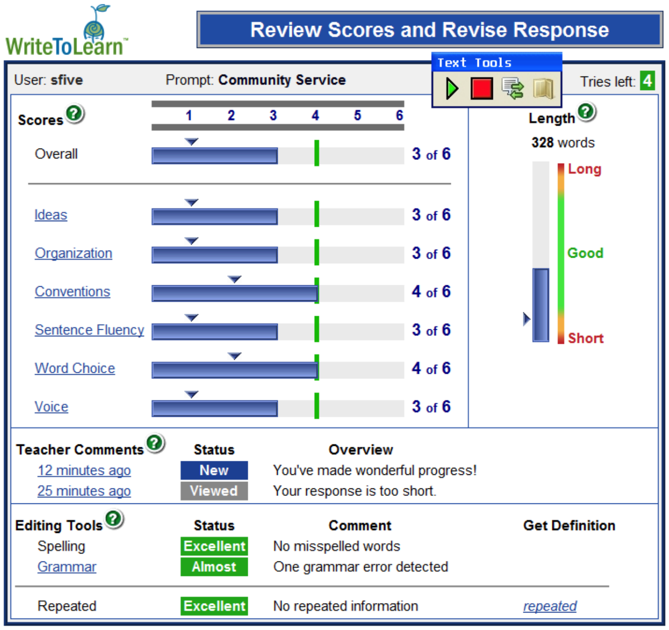
2 Implementation of automated scoring in a statewide assessment

In 2011, the State of South Dakota changed from a year-end summative writing assessment to a formative assessment that is run throughout the school year for grades 5, 7, and 10. The former assessment consisted of a summative 45-minute paper-pencil test that was administered each February. Students and teachers usually received results anywhere between 3-6 weeks after the completion of the test, precluding opportunities to use the results to work with students on areas needing improvement. The new state implementation is based on an automated formative writing assessment program to evaluate student writing. Students were required to write to at least three different exercises (writing prompts) each school year, allowing teachers to monitor progress and intervene where necessary to ensure that students were on track to meet learning goals [1].

2.1 Formative Writing Assessment

The formative writing assessment was implemented using WriteToLearn™. WriteToLearn is a web-based writing environment that provides exercises to write responses to narrative, expository, descriptive, and persuasive prompts as well as to read and write summaries of texts in order to build reading comprehension. Feedback is provided via overall and trait scores including "ideas, organization, conventions, word choice, and sentence fluency". In addition, grammar and spelling errors are flagged. Students are able to write, receive feedback and then revise and resubmit their improved essays. Tests of WriteToLearn have shown that it scores as reliably as human raters and results in significantly better comprehension and writing from two weeks of use[6].

Figure 1 below shows a portion of the WriteToLearn web interface indicating the system's scoring of a 12th grade persuasive prompt.



User: sfive

Prompt: Community Service

Tries left: 4

Scores	
Overall	3 of 6
Ideas	3 of 6
Organization	3 of 6
Conventions	4 of 6
Sentence Fluency	3 of 6
Word Choice	4 of 6
Voice	3 of 6

Length

328 words

Teacher Comments	Status	Overview
12 minutes ago	New	You've made wonderful progress!
25 minutes ago	Viewed	Your response is too short.

Editing Tools	Status	Comment	Get Definition
Spelling	Excellent	No misspelled words	
Grammar	Almost	One grammar error detected	
Repeated	Excellent	No repeated information	repeated

Figure 1. Essay Feedback Scoreboard. WriteToLearn provides students with an overall score as well as scores on six popular traits of writing. Passing scores are indicated by the green bars. Analysis of spelling, grammar, and redundancy (repeated information) is provided, as well as access to teacher comments. Clicking on individual traits provides more detailed explanations of how to improve those particular aspects of writing.

2.2 Algorithms for scoring writing

WriteToLearn's automated writing scoring is based on an implementation of the Intelligent Essay Assessor (IEA). IEA is trained to associate features extracted from each essay to scores that are assigned by human scorers. A machine learning-based approach determines the optimal set of features and the feature weights that best model the scores for each essay. From these associations, prompt and trait-specific scoring models are derived to predict the scores that the human scorers would assign to any new responses. Based on the prompt-specific models, new essays can be immediately scored by analysis of the features weighted according to the scoring models. In this paper, the focus is not on the actual algorithms or features that make up automated scoring which are described in detail elsewhere (see [5,7]). Instead, the focus is on how the automated scoring can be used to monitor learning across large sets of students and essay data.

3 Data Analysis

3.1 Data

During the Fall of 2011, teachers in the 5th, 7th and 10th grade assigned writing exercises to their students. The teachers were free to choose from among any of the pre-defined writing prompts in WriteToLearn for their writing assignments and incorporate them into their lesson plans. During that period, 21,137 students wrote to 72,051 assignments (an average of almost four assignments per student) with 107 different unique writing prompts assigned. For each assignment, students were required to submit at least one attempt per prompt, but could submit multiple revisions. By default, WriteToLearn allows a student to submit up to six drafts, but the teachers could set the default to be anywhere from one to 12. Overall, this setting resulted in 255,741 essays submitted and scored over the period of analysis. For each submission, students received feedback and scores on their overall essay quality, as well on six different writing traits, ideas, organization, conventions, word choice, sentence fluency and voice.

3.2 Analysis Goals

The goals of the analysis were to examine the data as a whole and were twofold:
1) Investigate the extent to which students improve their writing based on automated feedback
2) Examine the effects of revisions on the type of improvement in student essays

3.3 Number of student revisions

Figure 2 shows the distribution of revisions made by students. The greatest proportion of students submitted only a single draft. However, nearly as many students revised their essays five times (six submissions), which was the default maximum. The average number of revisions per student was 3.5 and the distribution clearly indicates that most students will take the opportunity to continue to modify their essays with feedback. A small proportion of students performed more than five revisions, attributable to the teacher increasing the default number of revisions.

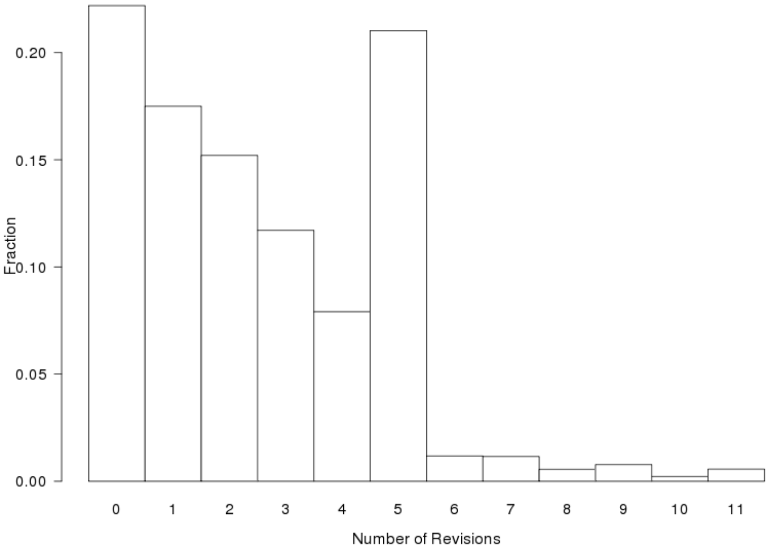


Figure 2. Distribution of revisions made by students.

3.4 Does writing and revising result in improved writing performance?

A critical question is whether the act of revising results in better writing, as measured by the automated scores using the IEA. A difference score was computed which compared the score the student received on their first submission to each subsequent submission. Figure 3 shows the score improvement (score on last attempt minus score on first attempt) for students who wrote multiple drafts. There is a clear trend indicating that with additional revisions, the students' score improved by greater amounts. With the typical five revisions, the average student score improved by almost one score point (out of a maximum of 6). The smoothness of the curve and small error bars are due to the large number of data points for each revision from 0 to 5.

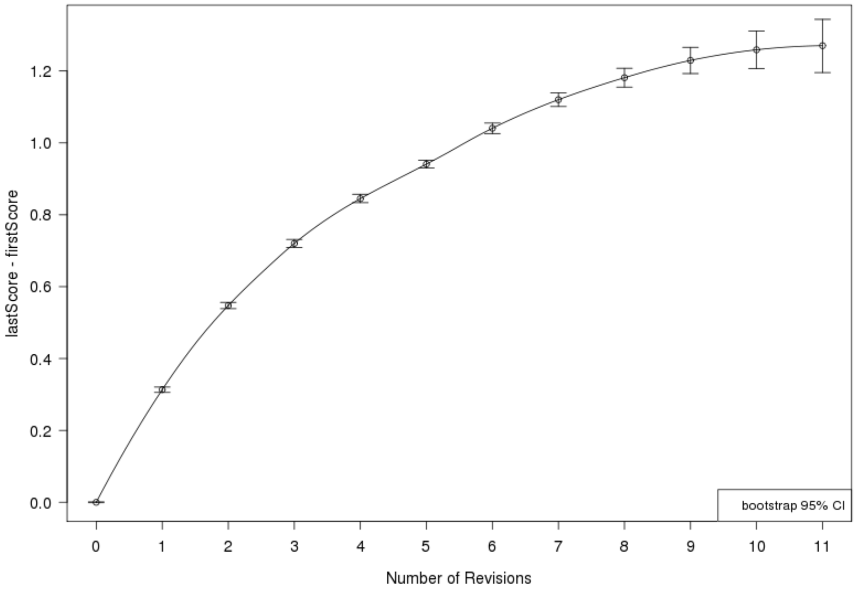


Figure 3. Overall score improvement with revisions

3.5 What aspects of writing improve with feedback?

For the set of prompts with scores on the six writing traits, we measured the change in scores from the first to last draft based on the number of revisions made by the students. Figure 4 shows the score improvement for the six traits as well as the overall score. Generally, we see greatest improvement in scores for ideas, voice and organization and less for sentence fluency and writing conventions. The results indicate that we see greatest improvement for the content of writing.

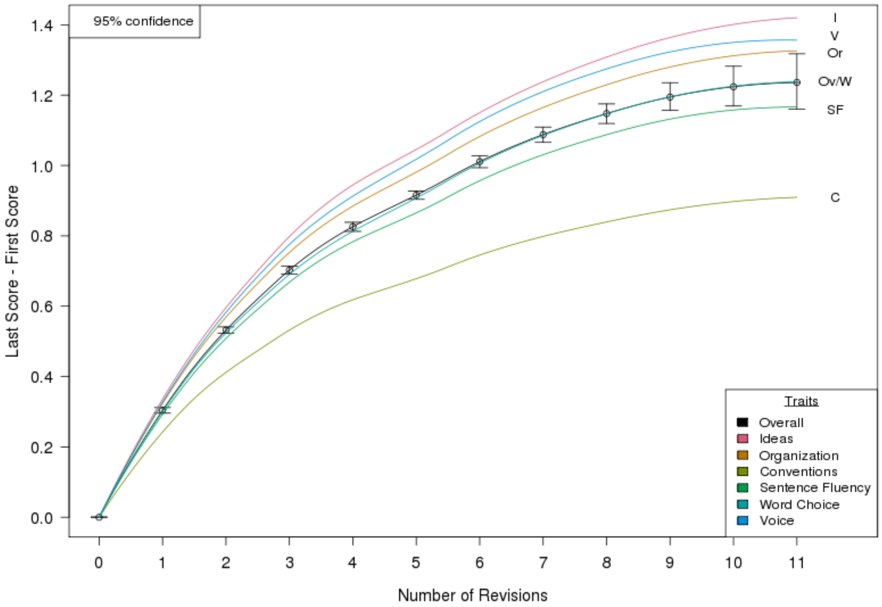


Figure 4. Change in writing scores for multiple writing traits across revisions

4 Implications

Large-scale formative writing assessment systems provide access to a rich set of data for analysis of performance and effects of feedback. Applying automated scoring of writing allows monitoring student learning as the students write and revise essays within these implementations. The overall results are not surprising; students improve with revisions. However, the approach provides means to examine the changes in learning and the effects of the feedback on their writing performance. For teachers, these insights can be used to help inform their instruction in real time. At a district and statewide level, this information can be used to help monitor progress in writing and track how policy and instructional changes may affect student performance in near-realtime. The results presented here provide an overview of a few of the analyses performed as part of an ongoing assessment administration. There is still much to analyze and ongoing work will help both improve methods of providing automated formative feedback as well as being able to give better information to teachers and administrators about their students' writing performance.

5. References

[1] Foltz, P. W., Lochbaum, K. E. & Rosenstein, M. (2011). Analysis of student ELA writing performance in a large scale implementation of formative assessment. Talk presented at the National Council on Measurement in Education. New Orleans, LA, April.

[2] Graham, S., & Perin, D. (2007). Writing next: Effective strategies to improve writing of adolescents in middle and high schools – A report to Carnegie Corporation of New York. Washington, DC: Alliance for Excellent Education.

[3] Graham, S., Harris, K., and Hebert, M. A. (2011). Informing writing: The benefits of formative assessment. A Carnegie Corporation Time to Act report. Washington, DC: Alliance for Excellent Education. Retrieved from http://carnegie.org/fileadmin/Media/Publications/InformingWriting.pdf

[4]Shermis, M. & Burstein, J. (2013). *Handbook of Automated Essay Evaluation*, M. Shermis & J. Burstein, (Eds.). Pp. 68-88. Routledge, NY. NY.

[5] 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.

[6] Landauer, T.K., Lochbaum, K.E., & Dooley, S. (2009). A new formative assessment technology for reading and writing. Theory into Practice, 48. 44-52

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