Automated feedback in a large-scale implementation of a formative writing system: Implications for improving student writing

Paper presentation at the American Educational Research Association Annual Meeting 2014

Peter W. Foltz
Mark Rosenstein
Nicholas Dronen
Scott Dooley

Pearson
Peter.Foltz@pearson.com

March 2014

Abstract

Practice can help improve students writing skills, most particularly when the students are supported with frequent feedback and taught strategies for planning, revising and editing their compositions. Formative systems incorporating automated writing scoring provide the opportunities for students to write, receive feedback, and then revise essays in a timely iterative cycle. This paper investigates the use of the formative writing tool WriteToLearn™ through mining student logs and essays in order to understand the use of the system and explore the extent to which students improve their writing based on the feedback from the system. In the implementation, students were given writing assignments and were able to write and revise essays. With each submission, the students received feedback on aspects of their writing including scores and instruction about different writing traits, redundancy, as well as grammar and spelling issues. The data collected included over a million student essays written in response to approximately 200 pre-defined prompts as well as a log record of all student actions, revisions and feedback given by the computer. Analyses examined the change in student performance over revisions of essays as well as the effects of different actions occurring within the system and the amount of time spent working on assignments. Implications are discussed for large-scale data analytics on writing assessment in order to understand the role of feedback in writing, to drive improvements in formative technology as well as to aid in designing better kinds of feedback and scaffolding for students to support the writing process.

Automated feedback in a large-scale implementation of a formative writing system: Implications for improving student writing

Introduction

It is a well known adage that in order to become a good writer, one needs to do a lot of writing. Along with sufficient practice however, good writing comes from receiving the right training and feedback. Meta-analyses of studies of formative writing (e.g., Graham, Harris, & Hebert, 2011; Graham & Hebert, 2010; Graham & Perin, 2007) have shown that supporting students with feedback and providing them instruction in strategies for planning, revising and editing their compositions can have strong effects on improving student writing. These studies further show that having teachers actively monitor a student's writing progress significantly improves student performance. However, scoring of writing can be time consuming, thereby limiting opportunities for students to receive timely feedback and limiting the teacher's ability to carefully monitor all students.

Automated scoring of writing, or Automated Essay Scoring (AES) provides the ability to analyze student writing and score writing instantly. Automated assessment of writing has become increasingly accepted with multiple systems available for implementing the scoring of writing (e.g., Shermis & Burstein, 2013). Studies of AES systems have shown that the scoring of such systems can be as accurate as human scorers (e.g., Burstein, Chodorow, & Leacock, 2004; Landauer, , Laham & Foltz, 2001; Shermis & Hamner, 2011), can score on multiple traits of writing (e.g., Foltz et al., 2013), can be used for feedback on content (Foltz, Gilliam, & Kendall, 2000), and can score short responses (e.g., Higgins et al., 2014; Foltz & Lochbaum, 2010)

While much focus has been placed on the accuracy of automated scoring, types of essays that can be scored and uses for summative scoring, AES also has wide applicability to formative writing. As a component of a formative tool, it can provide instantaneous feedback to students and support the teaching of writing strategies based on detecting the types of difficulties students encounter. For example, when incorporated into classroom instruction, students are able to write, submit, receive feedback and revise essays multiple times over a class period.

In a formative writing system, all student writing is performed electronically and automatically scored and recorded. Thus, there can be a record of all the student actions and all feedback they have 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 scrutinize 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 teaching effectiveness as reflected in student writing performance scores. At an even greater level of granularity, the data across multiple classes, schools, districts or states can be examined to examine changes in learning.

From a data analysis perspective, formative assessment of writing provides a rich data set to examine the changes in writing performance and the features of the system that influence that performance. Large-scale educational data analytics have examine a wide range of data from different types of learning environments including tutoring systems, gaming systems, and collaborative environments (e.g., Romero, Ventura, Pechenizkiy, & Baker, 2010). Similarly formative writing systems present an opportunity for large-scale analytics on writing. Prior work (Foltz, Lochbaum, & Rosenstein, 2011) has examined changes in student writing on a smaller data set from a state-wide implementation of an automated formative writing system. The results demonstrated that students can improve over revisions in writing, with greater improvement shown on writing aspects such as content and organization.

The goal of this paper was to examine the changes in writing performance and the features of the system that promote performance through an analysis of subsets of data from large-scale implementations of the WriteToLearn formative writing system. When a system is deployed in multiple educational settings across a wide range of grade levels and contexts from rural to urban schools an enormous space of use patterns emerge.. Using data from over a million student draft essay submissions along with a similar order of magnitude log of student actions, there are a wide range of different types of analyses that can be performed. These include measuring improvements in student writing across drafts, better understanding of the use of different writing tools by the students, discerning what aspects of writing are more apt to be improved through automated feedback and the effects of different types of feedback (e.g., grammar feedback vs. feedback on writing traits such as ideas, organization, or word choice) on student writing performance. This paper illustrates some examples of the kinds of information that is made available from analysis of components of the formative writing process.

Method

WriteToLearn

The formative writing assessment system used for the analyses was WriteToLearn™. WriteToLearn is a web-based writing environment that provides students with 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. 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 such as "ideas, organization, conventions, word choice, and sentence fluency". Supplemental educational material can also be viewed by the student to help them with understanding the feedback, as well as indicating approaches to improve their writing. In addition, grammar and spelling errors are flagged. Figure 1 below shows a portion of the system's interface, in this case illustrating the scoring feedback resulting from a submission to a 12^{th} grade persuasive prompt. Evaluations of WriteToLearn have shown significantly better reading comprehension

and writing skill resulting from two weeks of use (Landauer, Lochbaum, & Dooley, 2009) as well as validating the system scores being as reliable as human raters.

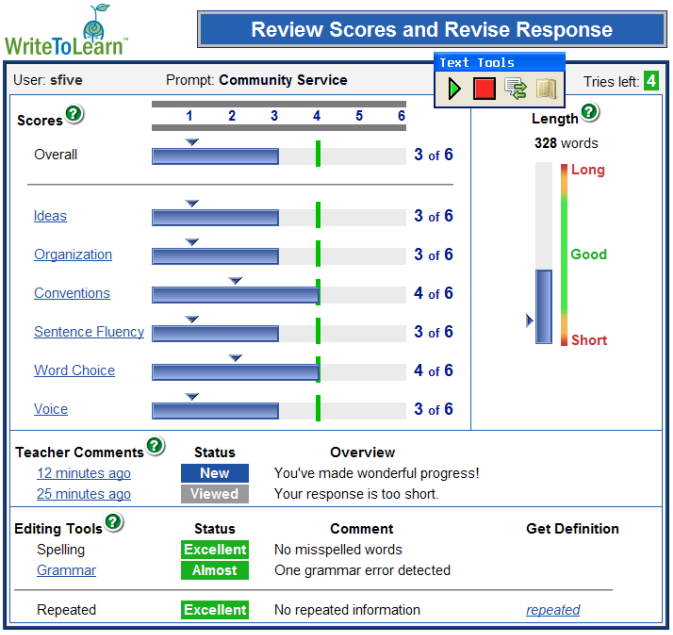


Figure 1. Essay Feedback Scoreboard. WriteToLearn provides students with an overall score as well as scores on six popular traits of writing. Passing scores are shown by the green bars. Analysis of spelling, grammar, and redundancy is provided. Clicking on individual traits provides more detailed explanations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

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 extracted features from each essay to scores that are assigned by human scorers. A machine learning-based approach is used to determine the optimal set of features and the weights for each of the features to best model the scores for each essay. From these comparisons, a prompt and trait-specific scoring model is derived to predict the scores that the same scorers would assign to any new responses. Based on this scoring model, new essays can be immediately scored by analysis of the features weighted according to the scoring model. The focus in this paper is not on the actual algorithms or features that make up the scoring as those have been described in detail elsewhere (see Landauer, Laham, & Foltz, 2001; Foltz et al., 2013). Instead, the focus is how the trail left by automated scoring and student actions can be used to monitor learning across large sets of essay data.

Data

The data collected included student essays as well as 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, so there was a record of each draft submitted, but not individual keystroke level information from the editing process as each essay draft was created. The data was compiled from students in

multiple regions of the United States who generated more than a million essays written to approximately 200 pre-defined prompts. It should be noted that no human scoring was performed on the essays; all essay scores were generated by automated scoring.

Results

Levels of Granularity

We can view and understand student use and learning at multiple levels of granularity. In examining the time-scale of use, analyses can range from looking across multiple years of a student's progress down to as small a level as examining individual student actions. Some students have interacted with WriteToLearn for periods of up to four years and therefore there is a record of changes in individual writing across a significant span of their grade school development. Even at the level of a school year, we can examine changes in student writing quality from one prompt to the next. Of equal importance we can track their learning within a single assignment, and since assignments often span multiple class periods, we can track their learning even within a given class period, which we call a session. Finally, at the session level we can interpret their actions, such as use of help or obtaining feedback to be used to modify and individualize their instruction. It should be noted that in this paper, we have analyzed relevant subsamples of the data, so not all levels of granularity are fully described or analyzed

Number of revisions made by students per prompt

One of the proposed advantages of automated formative writing is that it supports a rapid cycle of write, submit, receive feedback and revise. Thus, it is critical to examine how often students do submit essays and revise. By default, the maximum number of submissions in WriteToLearn is six, though this limit can be modified by the teacher assigning a prompt. Figure 2 shows the distribution of submissions made by students per writing prompt. The two modes of the distribution show that nearly equal proportions of students submit a single attempt as submit the full six submissions. The distribution clearly indicates that most students will take the opportunity to continue to modify their essays with feedback. A small proportion of students submitted more than six revisions, which indicates that the teacher increased the default number of revisions. These results are similar to the distribution found previously of a smaller sample of about a quarter of a million of essays (Foltz et al., 2011). The results indicate that students are taking advantage of revising essays and resubmitting for feedback. This plot raises additional research questions. For example, we can then investigate the patterns of submissions for feedback as well as better understand the use of the system for students who only submitted once and did not to avail themselves to feedback.

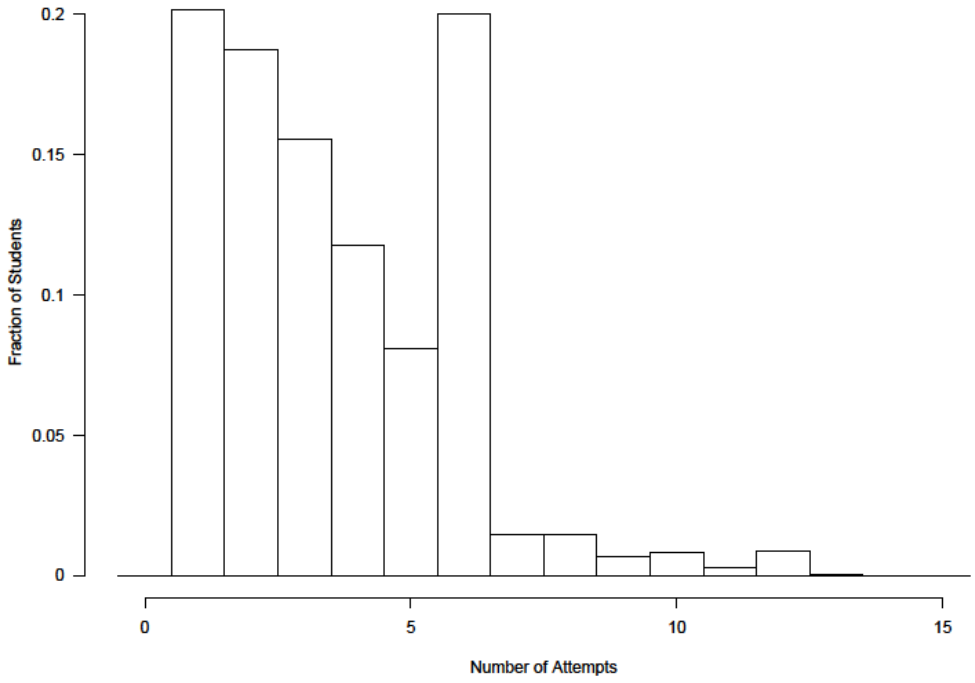


Figure 2. Number of student submissions (attempts) as a proportion of the total students.

Time spent on individual writing assignments.

Along with understanding the number of times an essay is revised, it is equally important to understand the temporal pattern of student use. One measure of temporal use is to examine the time between the initial submission and the final submission per student on a single writing prompt. Figure 3 shows assignments lasting up to seven weeks in duration. It was necessary to truncate the first bar, which essentially covers a single class period, since it represented 48% of the students. Thus, the vast majority of student work on an assignment that encompasses a single class period. The second largest spike, which contains approximately 11% of the students are assignments that are finished over a little more than 1 day (two class periods). There are smaller spikes up to a week, and a week is the next largest proportion after 1 or 2 class periods, but interestingly, there is a continued pattern of final assignments ending after 2, 3 and even beyond 7 weeks, though the plot truncates at 7 weeks.

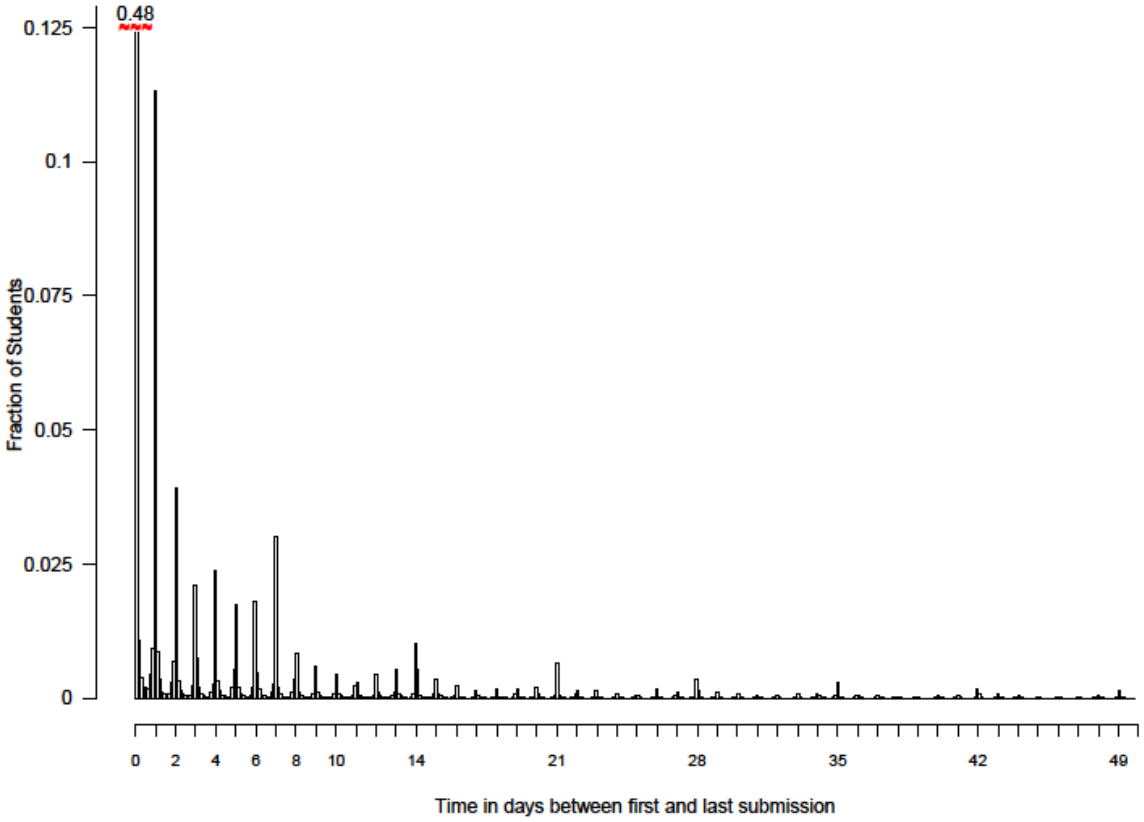


Figure 3. Time between first and last submission per assignment

Noting this behavior, we can ask what effect increasing the length of time of the assignment has in terms of increase in score between first and last attempt. This effect provides some indication whether students improve their writing over longer revision periods. The next plot indicates the impact on the overall score change as the length of time spent on the assignment increases.

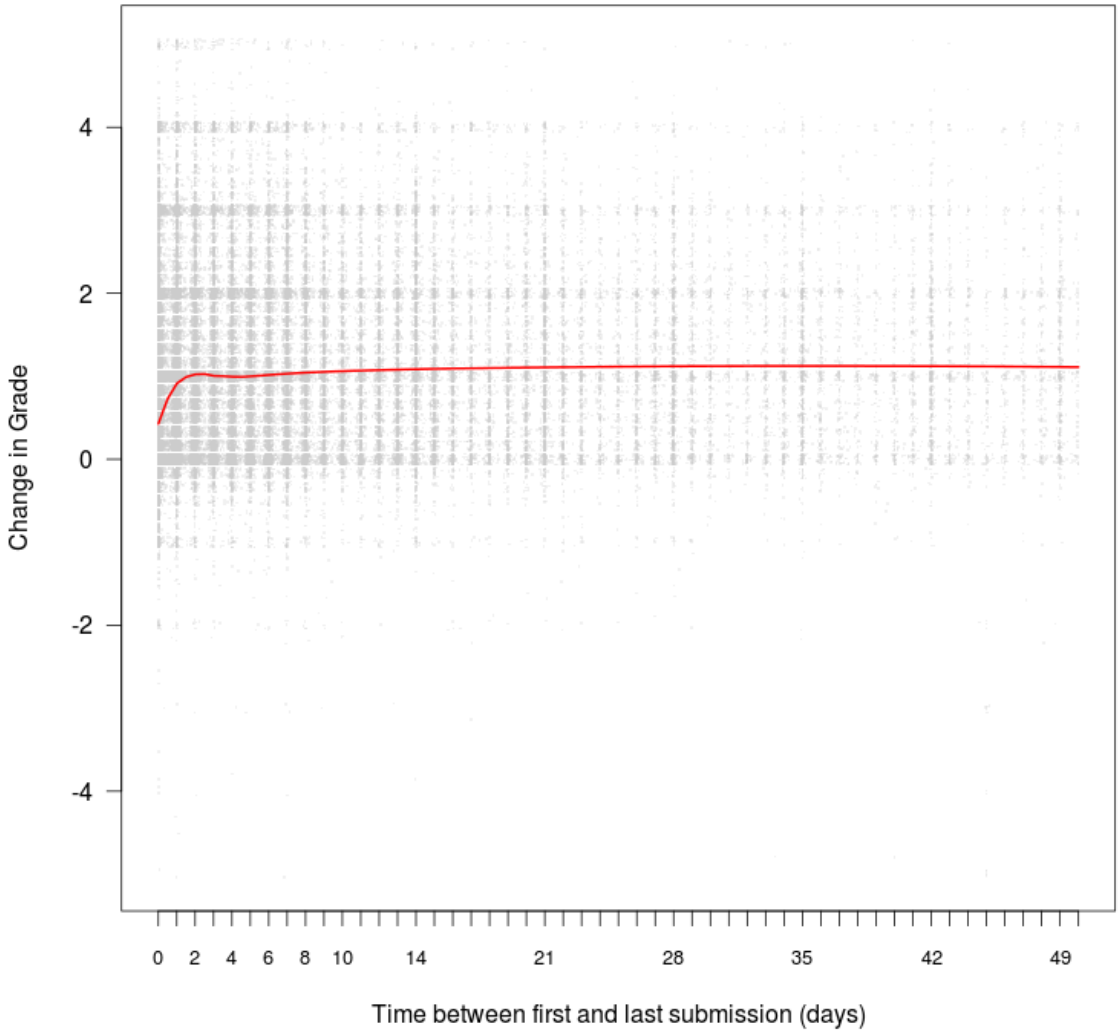


Figure 4. Change in grade relative to time on assignment

Figure 4 indicates the change in score as grey points (overwriting due to the large number of students), and a locally weighted regression line in red summarizing the data. We see that the vast majority of students improve (most of the mass of grey is above the y=0 line), with an average improvement of about one score point (All scores were on a 1-6 scale). Only a small number of students in the first few days actually did not improve their scores. The regression curve hints that most of the improvement is seen in students that work on the assignment over the first few days, and students working on the assignment longer don't seem to gain additional benefit. Of course these data are not experimentally controlled, so there may be other confounding factors that will need to be investigated, such as students forgetting to submit assignments and then not making changes up to much later deadlines. The results though suggest that keeping an assignment open for weeks versus moving on to another prompt may not be the best strategy.

As indicated in the plot from Figure 4, there is a wide range of time spent on a given prompt. We can also examine the distribution of time spent before each submission (after the first). Figure 5 shows that this distribution also has a long tail, but since most of the assignments are for a single class period, the following histogram is for submissions less than 40 minutes.

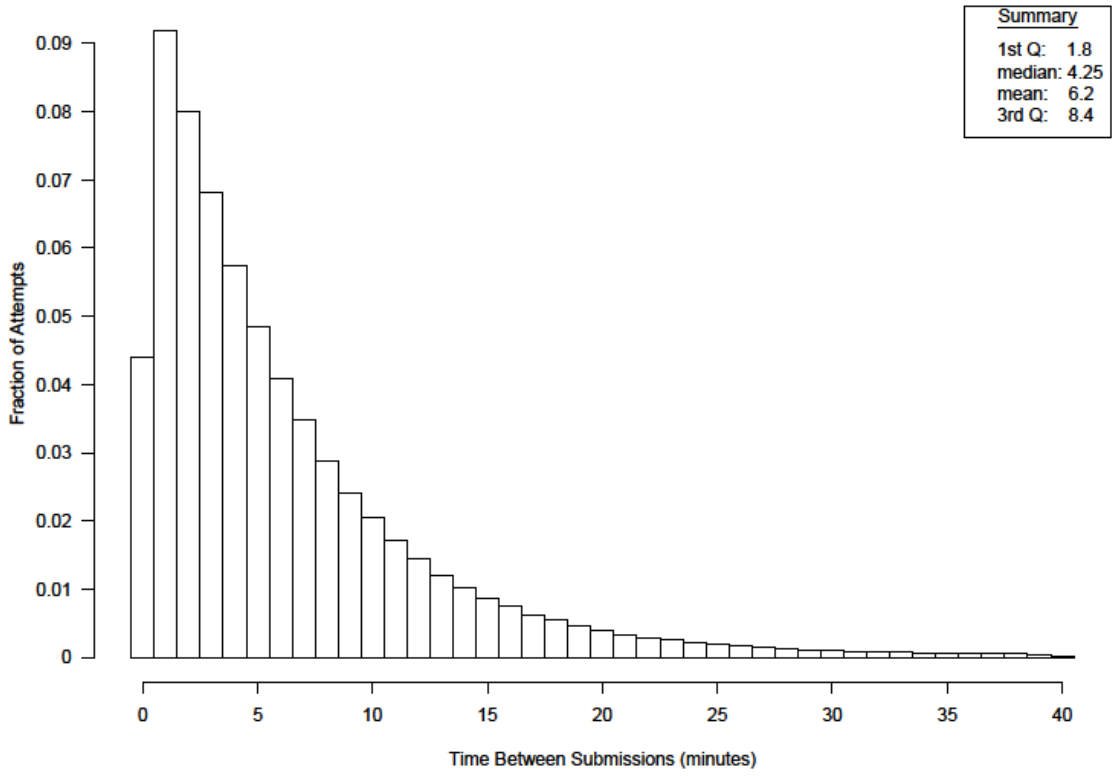


Figure 5. Proportion of submissions as a function of time between submissions

The results indicate a mean time between essay submissions of 6.2 minutes and a median of 4.3 minutes. It is interesting to note that a large proportion of students are resubmitting essays in under a minute, indicating likely very minimal revising on the part of the student. In order to understand the impact of the amount of time spent between revisions, we can look at the change in grade versus the time spent revising. Figure 6 illustrates this, with gray indicating individual student points, and a locally weighted regression in red describes the trend.

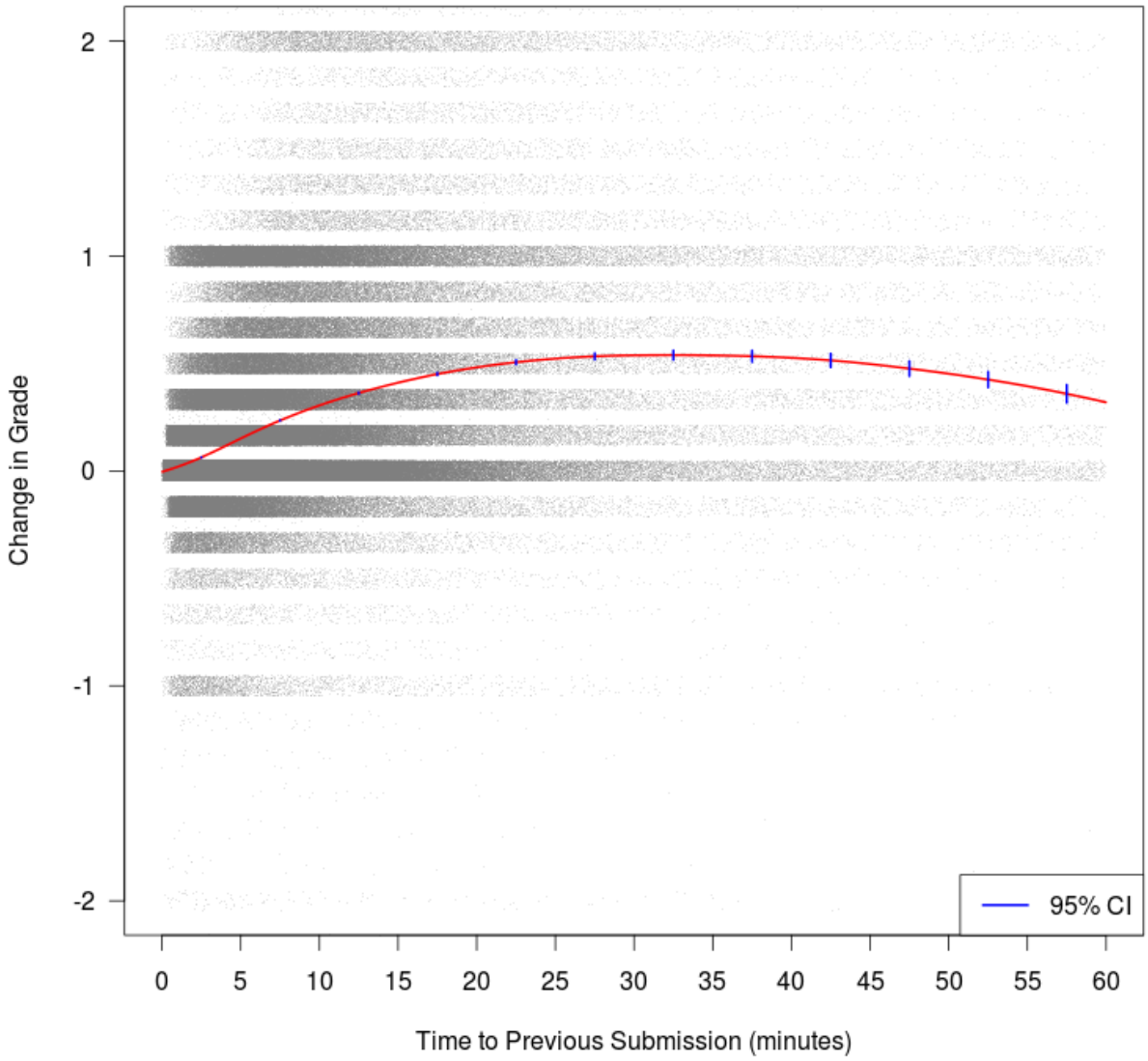


Figure 6. Change in grade based on time from previous submission.

The results indicate 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. It seems clear that a strategy of making a small number of changes and then quickly resubmitting does not suffice for writing improvement. Again Figures 5 and 6 raise research questions about the strategies students use in their submissions. We would like to better understand the differences between the revisions that were submitted very quickly versus those where students allocated

more time before submitting. This requires analysis at the action level of granularity.

Session Analysis – Actions

The results above show that students do revise their essays and that the time pattern of essay revision suggests that the greatest improvement in writing from one revision to the next is around 15 to 40 minutes. The results don't indicate though what is occurring during that time. Thus, the next analyses examine actions that were taken by students within a session. Actions in the session can include such aspects as logging in/out, submitting an essay for assessment, performing spell or grammar checks, requesting additional formative information about writing (such as seeing scoring rubrics, how to improve writing on different traits, etc.). Individual keystroke level editing actions between essay submissions or saves were not recorded. A session is defined as the time between when a student logs in and when the student logs out or the student is automatically logged out after an idle period timeout. Most typically during a session a student will work on writing for one prompt, although a student could potentially work on more than one prompt if more were assigned by the teacher, and which occurs in approximately 10% of sessions.

If we examine the time spent on individual sessions, we see that, while many sessions are very short (less than five minutes), the bulk of the writing sessions are class period length of about 25 to 50 minutes, with some students spending upwards of two hours as shown in Figure 7.

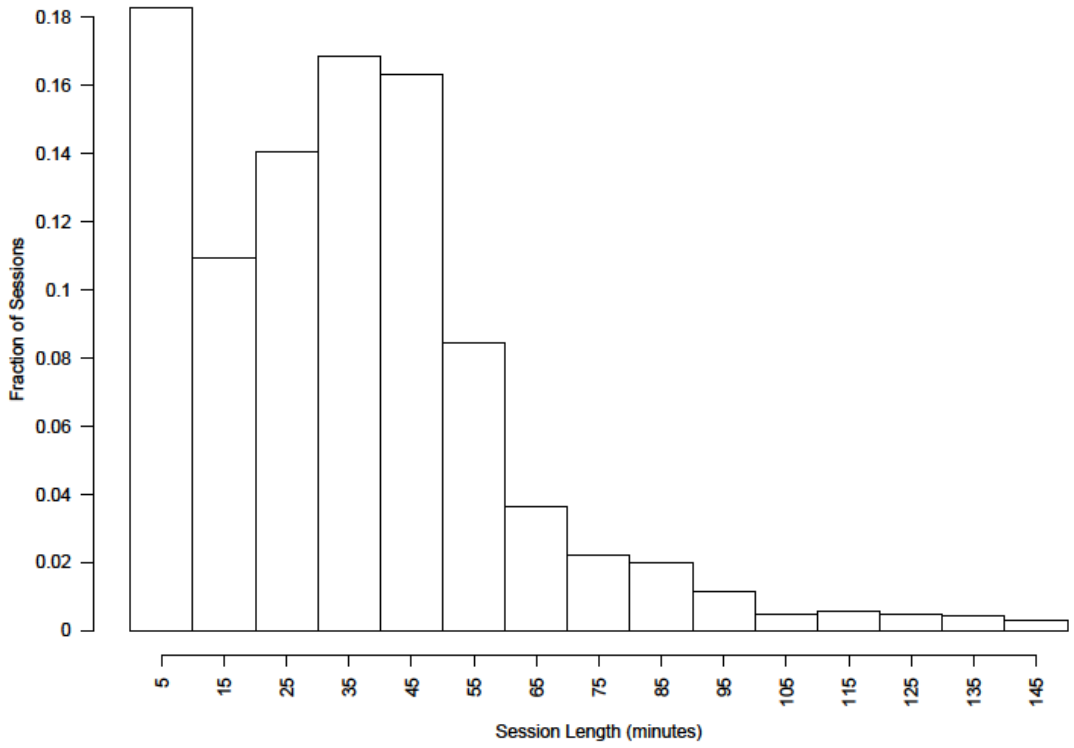


Figure 7. Proportion of sessions related to their length in minutes

Along with determining the length of time spent in a session, we can also examine the number and types of actions that were performed during a session. The number of actions provides some indication of the how active the student was in using the different features within the system. Figure 8 shows the distribution of the number of actions performed by students during a session. The results indicate that beyond editing the text, students are performing a number of actions, with the large number of students performing at least 10 actions.

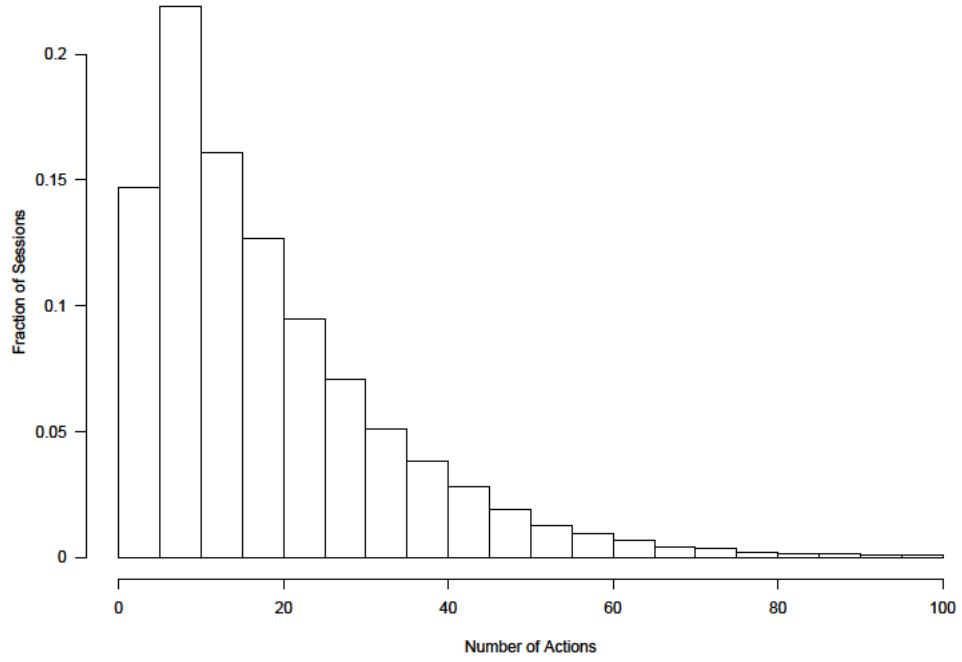
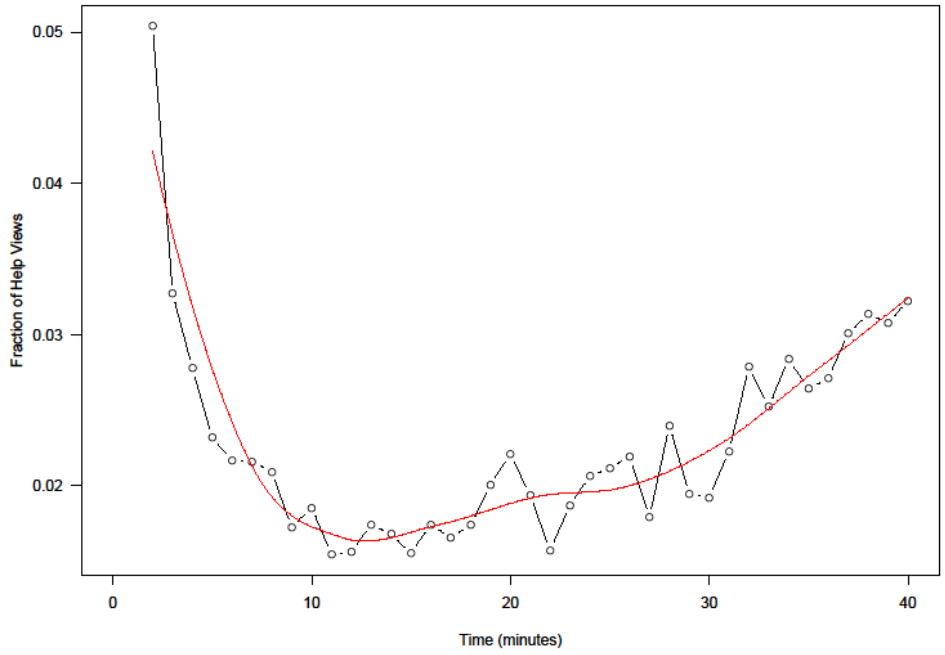


Figure 8. Distribution of the proportion of actions taken by students in a single session

The number of actions provides just an indication of activity, but not the type of activity being performed. We can therefore analyze the types of actions performed to gain some insight into the use of features. In Figure 9 we illustrate the proportion of times that students requested additional writing guidance (called help views) as a function of time in their session. It is interesting to note that the highest proportion of writing guidance is used very early on in the session. Requests for writing guidance then decreases quickly, likely indicating times where students are involved in writing, but then increases later in the session again, most likely when they have received feedback and are working on revisions. The results provide some indications of how writing guidance is used in the writing process. They further indicate potential areas that the software could be improved. For example, the software could potentially provide easier access to writing guidance during the writing process.



.
Figure 9. Use of additional writing guidance as a function of time in a session

Patterns of usage across actions

We can further understand the usage of the system by examining the temporal flow of multiple actions within a student session. Such an analysis can provide an indication of which actions are being performed and can then be related back to whether the pattern of actions are indicative of successful performance. While a majority of the action patterns show good progression and use of the different features of the writing system, one can also detect patterns that may be suboptimal to learning.

Figure 10 shows the pattern of action from a student over a 33 minute session in which the essay score did not change from one revision to the next. The student's action timeline is spread over three lines, the top representing the first 10 minutes of the session, the middle 10 to 20 minutes, and the bottom row the last 13 minutes of the session. During the session, the student performs 73 grammar and spell checks, but only requests feedback from the computer twice. The student further only looks up additional writing guidance once, which occurs near the end of the writing process. While the above pattern indicates a student who is requesting only limited feedback from the computer, Figure 11 shows the action path for a sample of 13 students, each who submitted for feedback twice or more in a row, with less than approximately 2 minutes between submissions (shown in yellow) shown in the complete context of the action sequence of their session. These are cases where students likely did not make major revisions to their essays and likely would have benefited more from guidance on further revising their essays before resubmitting.

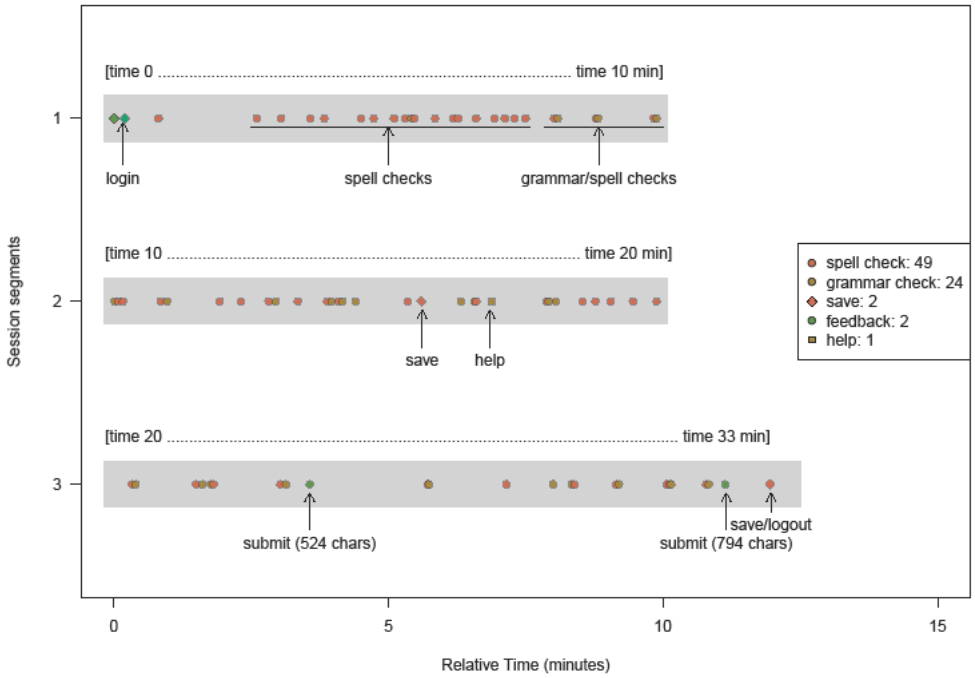


Figure 10. Actions performed by an individual student

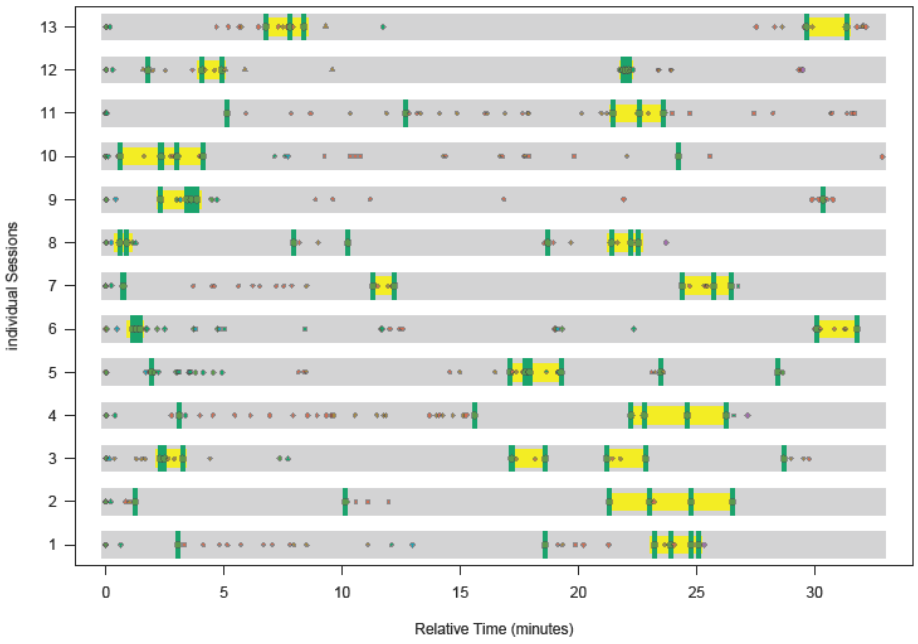


Figure 11. Actions performed by 13 students who resubmitted essays in less than two minutes

In each of these cases these patterns can be readily detectible and allow providing more scaffolded support to the student on the writing process. An analysis based on these patterns provides the potential to support different types of writing work flows that may align with different types of teaching or for writing of different essay types. For example, writing an essay in response to a text may have a workflow pattern where students go back to the original text, annotate (e.g., locate claims and evidence) and then incorporate them into the structure for writing, whereas an expository essay may require patterns of more revisions and support structures for descriptions or comparisons.

Measuring changes in essays across revisions

Logs of student actions therefore provide a rich representation of the choices students made and their paths through the learning environment. Essays, on the other hand, provide one of the richest sources of information about the students' knowledge and writing abilities as well as what they have learned. Tracking changes in the content, structure, words, or characters of essays from one revision to the next can allow measures of how student's knowledge has changed. One fundamental measure of change in writing is the edit distance between two drafts. Edit distance quantifies the amount of change between two strings (e.g., revisions of the same essay) based on the number of insertions, deletions, and substitutions required to transform one string into the other. In Figure 12, we illustrate the analysis of edit distance between two successive revisions for 100 essays. It is evident that from one revision to the next, the majority of changes are insertions (additions of new information). A good number of the revisions though are very minor changes, resulting in modifying fewer than 75 characters in an essay (about 12 words). By tracking these kinds of changes, one could potentially provide additional scaffolding to students who have made minimal changes.

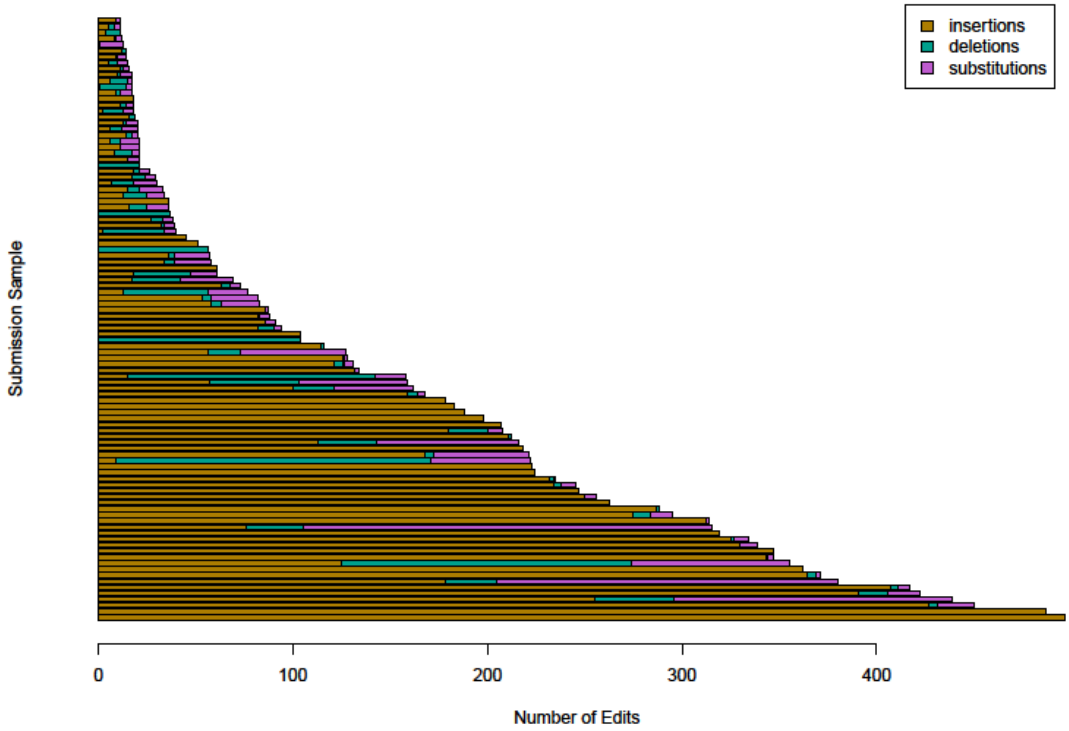


Figure 12. Edit distance for a sample of student revisions.

Edit distance measures provide just a basic, initial characterization of the changes without a complete understanding of what changed. Computational language tools can provide the potential to perform much finer characterization than edit distance. For example, by using natural language processing models including statistical and n-gram models, semantic analyses, and syntactic parsing, one can analyze aspects such as changes in the structure of the text, inclusion of new, relevant semantic content, changes in syntactic structures, and revisions to the coherence of the essay. Additional work is ongoing to provide this deeper level of characterization in writing over revisions in order to better understand and characterize the revision process.

Conclusions

Large-scale implementations of formative writing provide rich sets of data for analysis of performance and effects of feedback. Applying 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 track the effect of use of the system.

The overall results are not surprising; students generally improve with revisions and feedback. However, the approach described here provides means to examine the changes in learning and the effects of the feedback on writing performance. The results show the amount of time spent that produced the greatest gain. Very quick revisions to essays did not tend to produce better essays. Instead, greatest writing growth from one revision to the next occurred with about 20 to 40 minutes of intervening time, most presumably spent writing and learning.

The results presented here provide an overview of a few of the analyses performed as part of ongoing investigation into the use of data-mining for formative assessment. In addition, there are a number of limitations in analyzing data based on very large data sets. All improvement was measured based on the automatically generated scores, since there is no way to validate millions of essays by human scorers. While there have been many studies which show that automated essay scoring closely agrees with human scores, it does not rule out the possibility of the system driving students to do better at what the system is good at scoring. WriteToLearn is used in many varied contexts, including for homework and classwork. A teacher may assign work differentially, sometimes having students write quickly or focusing on different aspects of writing such as just improving organization. Therefore there is no control over the contexts in which students are creating data. However, by analyzing very large amounts of data, some of these varied contexts may be averaged out. In addition, further work could detect some of the different patterns of usage that may indicate these different contexts, which would allow better targeting of feedback.

There is still much to analyze. Ongoing work is focusing on more formal modeling of the effects of actions on student performance and of the action sequences. The work will help improve methods of providing automated formative feedback, provide better information to students about how to improve their writing and help teachers and administrators better understand and use information about their student writing performance. For the teachers this information 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.

References

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

Foltz, P. W., Gilliam, S. & Kendall, S. (2000). Supporting content-based feedback in online writing evaluation with LSA. *Interactive Learning Environments, 8*(2), pp. 111-129.

Foltz, P. W. & Lochbaum, K. E. (2010). AI scoring MSA science. Paper presented at the meeting of the *Maryland Assessment Group*. Ocean City, MD.

Foltz, P. W., Lochbaum, K. E. & Rosenstein, M. (2011). Analysis of student writing for a large scale implementation of formative assessment. Paper presented at *the National Council for Measurement in Education.* New Orleans, LA.

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.

Graham, S., Harris, K. R., & Hebert, M. (2011). Informing Writing: The Benefits of Formative Assessment. A Report from Carnegie Corporation of New York. *Carnegie Corporation of New York*.

Graham, S., & Hebert, M. (2010). *Writing to read: Evidence for how writing can improve reading: A report from Carnegie Corporation of New York*: Carnegie Corporation of New York.

Graham, S., & Perin, D. (2007). A meta-analysis of writing instruction for adolescent students. *Journal of Educational Psychology, 99*, 445-476.

Higgins, D., Brew, C., Hellman, M., Ziai, R., Chen, L., Cahill, A., Flor, M., Madnani, N., Tetreault, J., Blanchard, D., Napalitano, D., Lee, C. M., & Blackmore, J. (2014) Is getting the right answer just about choosing the right words? The role of syntactically-informed features in short answer scoring. *arXiv*: 1404.0801v2.

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

Landauer, T., Lochbaum, K., & Dooley, S. (2009). A New Formative Assessment Technology for Reading and Writing. *Theory into Practice, 48*(1).

Romero, C., Ventura, S., Pechenizkiy, M and Baker, R. (2010). *Handbook of Educational Data Mining*. New York: Taylor & Francis, 2010.

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.

Shermis, M.D., & Burstein, J. (2013). *Handbook of Automated Essay Evaluation: Current Applications and Future Directions*. New York: Routledge.