Analysis of student ELA writing performance for a large scale implementation of formative assessment

Peter W. Foltz
Karen E. Lochbaum
Mark Rosenstein

Pearson Knowledge Technologies

ALWAYS LEARNING

PEARSON

Automated Formative Writing Assessment

Learning to read and write well comes from doing a lot of reading and writing

- But human scoring is time consuming, limiting opportunities for providing timely feedback

Automated Assessment of Writing

- Provides more opportunities to practice reading and writing skills with feedback
- Provides specific, immediate and individualized feedback
  - -Addresses content of writing as well as surface-level features
- Embeds assessments within natural performance tasks
  - Promotes abilities that will be used in and out of the classroom.
- Permits monitoring of performance changes in individual as well as large groups of students

PEARSON

Many studies out there show that time on task is a very strong predictor of performance gains in reading and writing.   Particularly for writing, it's not just writing, but it is also getting feedback on writing .

In this talk, I will describe an example of applying automated scoring technology in formative writing assessment . I will discuss some analyses of the results of its use in a large scale implementation and implications.

Automated Formative assessment of writing can provide extensive opportunities for students to write and receive specific and immediate individualized feedback.   It further allows teachers to be more effective at monitoring the performance of students and adjust their teaching as needed.

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 promote performance.

The goal of the study was to examine student learning through measuring improvements in performance in a large set of data on formative writing

WriteToLearn

Online tool for building writing skills and developing reading comprehension

- Writing instruction through practice
- Reading comprehension through summarization
- Immediate, automated evaluation with targeted feedback
  - Six traits of writing
  - Summary quality and missing information
  - Grammar, spelling, redundancy, off-topic sentences, …

Studies of WriteToLearn components compared to control groups

- Significantly better comprehension and writing from two weeks of use (Wade-Stein & Kintsch, 2004)
- Increased content scores compared to controls (d=.9) (Franzke et al., 2005)
- Improved gist skills on standardized comprehension test (d=.42)
- Scores as reliably as human raters

PEARSON

WriteToLearn is a web-based tool that integrates practice and assessment incorporating reading comprehension with expository writing.

Results of prior studies indicate that using automated formative writing practice improves writing skills. It is not just the act of writing that results in improvement but getting the directed feedback that provides the greatest gains.   However, most of the studies have been small-scale using a few classrooms.   The present study examines a very large data set to examine if we see similar results in improved writing performance.

Reliability of WriteToLearn Essay Scoring Component

- 99 diverse prompts; 4th-12th grade students
- Scoring developed using essays with scores by operational readers of a major testing companies.
- Trained on essays, tested on others

Measure	Automated Scoring to human raters (min, mean, max)	Human raters to human raters (min, mean, max)
Correlation	.76 **.88** .95	.74 **.86** .95
Exact score agreement	50% **63%** 81%	43% **63%** 87%
Exact + adjacent agreement	91% **98%** 100%	87% **98%** 100%

PEARSON

The reliability of the WriteToLearn essay scoring has been tested over a wide range of prompts and grade levels.   Generally, we see levels comparable to human levels of reliability.

The essay prompts include persuasive, how-to, response to literature, descriptive, expository, autobiographical, narrative, workplace writing, comparison and contrast, problem and solution.

The focus of today's talk is not on testing the reliability, but on assuming that we have reasonable reliability, which these results indicate that we do, to in order to examine changes in student writing performance.

The question I'm addressing today is not "is this as reliable as a human?" (which it is), but "do we have a tool that helps students learn?"

South Dakota implementation of WriteToLearn

Formative Assessment replaced year-end summative writing assessment
- State requirement for writing
- Grades 5,7,10
- Provides teachers with opportunity to use results to work with students on areas needing improvement

Evaluate student writing at least three times per school year
- Teachers choose prompts, decided how to incorporate into their lessons and set required level of student achievement
  - Pre-defined prompts, summaries, teacher created prompts,
 - Students required to submit at least one attempt per prompt

South Dakota has changed from a year-end summative writing assessment to a formative assessment that is run through the school year. The old test was 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.

"When student writing was only evaluated once at the end of the school year, teachers didn't have the opportunity to use the results to work with their students on areas needing improvement".

"Using WriteToLearn to evaluate student writing three times each school year, teachers can monitor progress, intervene where necessary and ensure that all learners are on track to meet learning goals. "

It can also help teachers teach writing.
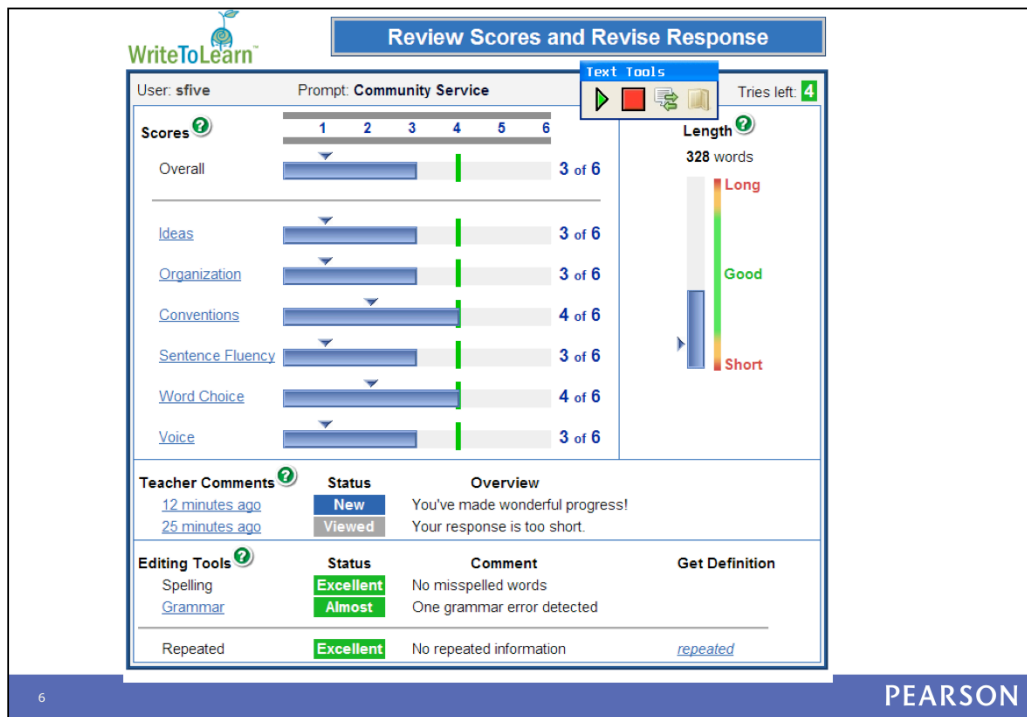
An example of the WriteToLearn student interface.   Students write an essay and receive feedback on 6 traits of writing, indications of how they performed on their last draft as well as comments from their teachers on drafts. They further have editing tools to help improve spelling, grammar and redundancy in their writing.

Analysis of South Dakota Writing data

Goals
- Investigate the extent to which students improve their writing based on automated feedback
- Examine effects of revisions on the type of improvement in student essays

Essay prompt data

21,137 students wrote on 72,051 assignments against 107 pre-defined prompts

255,741 scored submissions

Average of 3.5 drafts per student per prompt

PEARSON

The goal of the data analysis was to investigate the extent to which students improve their writing based on automated feedback and the effects of students making multiple revisions on the quality of their essays.

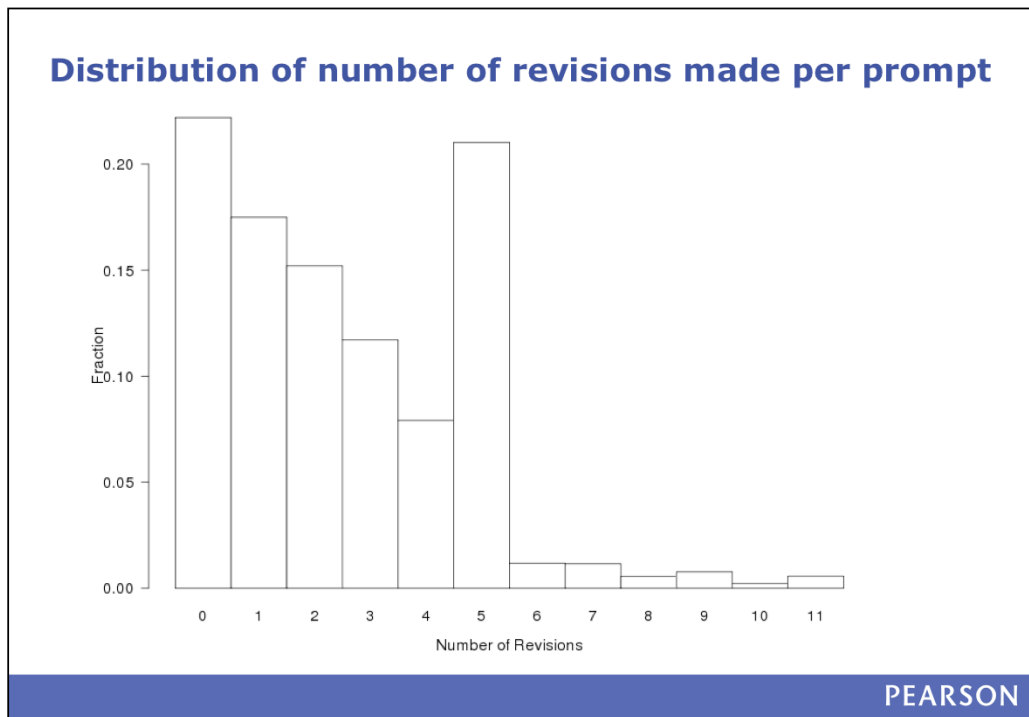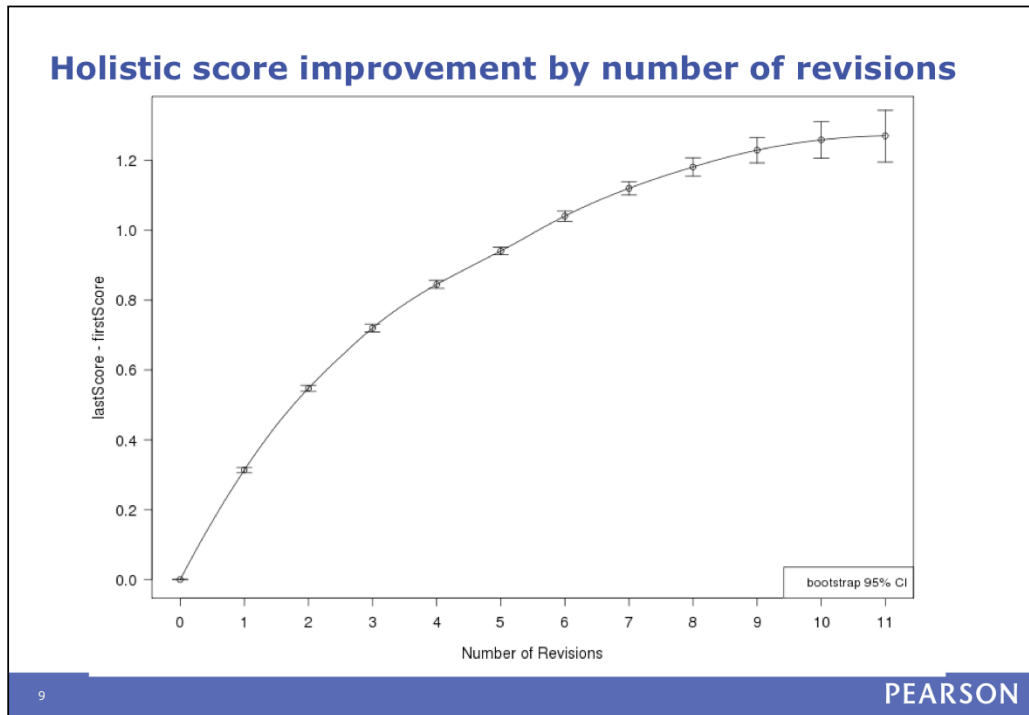**Distribution of number of revisions made per prompt**

Figure 1 shows the distribution of the number of revisions performed by students on each prompt. By default, the number of allowed attempts is 6, but can be set as high as 12 by the teacher.

Holistic score improvement by number of revisions

Does writing and revising drafts result in improved writing performance?   This figure shows the score improvement (score on last attempt minus score on first attempt) for students who wrote multiple drafts. It shows that with more 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).

Score improvement for different writing traits based on number of revisions
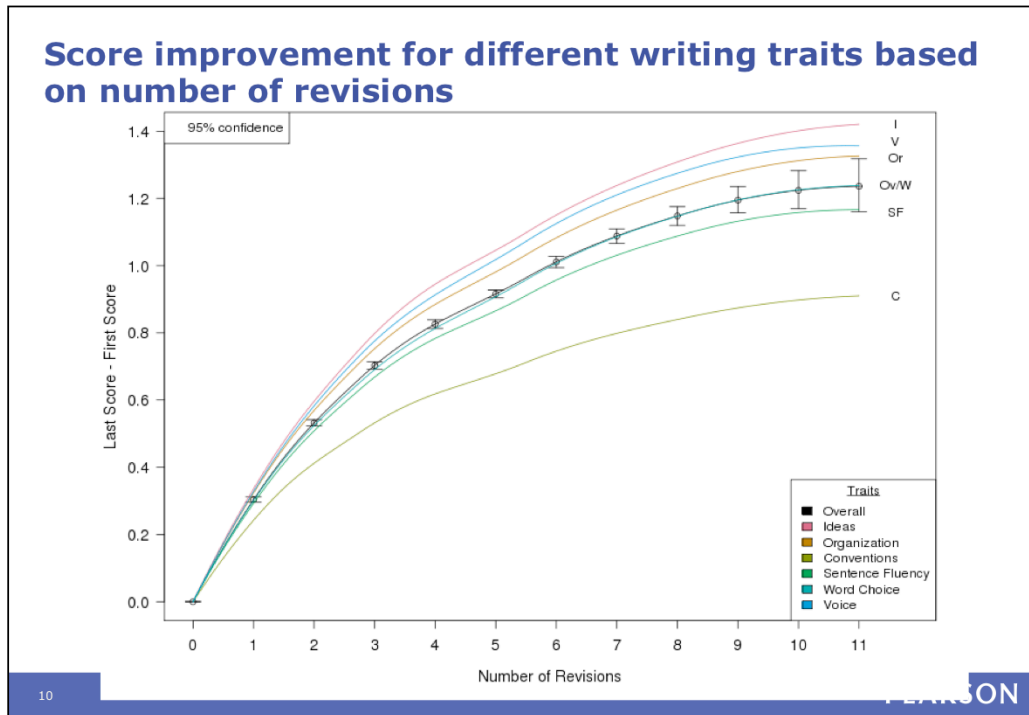
What aspects of writing improve with WriteToLearn feedback? For prompts that scored students on the six traits of writing, we measured the change in scores from the first to last draft based on the number of revisions made by the students. Figure 3 shows the score improvement for the six traits of writing 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.

Further work will examine the degree to which these differences may be due to the sensitivity of the measurement of the traits, features used that cause changes in performance as well as differences across grade levels.

Conclusions

Implications
- Online formative assessment provides rich set of data for analysis of performance and effects of feedback
- Evidence for strong effects of improvement on writing quality through process of revisions
- Automated assessment of writing can track changes in performance from individual to classroom to state-wide
- Integrated into the classroom, it helps foster training teachers in providing writing feedback

Ongoing work: Data mining in formative assessment
- Student performance changes over the year
- Changes in performance for writing traits based on student level and type of feedback received
- Comparisons to human scoring, external measures

Students improve with revisions. That's not a surprising finding. However, we want to be able to see that it is happening as they write and revise and the effects of the feedback on their writing performance. For the teachers this can be used to help inform their instruction in real time. At a district and statewide level, this can be used to help monitor progress in writing.   The results presented here are provide an overview of a few of the results of a large amount of data. 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 users about their student writing performance.