Integrating Product and Process Data in an Online Automated Writing Evaluation System

Chaitanya Ramineni
Educational Testing Service
Princeton
NJ, 08541
01+609-734-5403
cramineni@ets.org

Tiago Calico
University of Maryland
College Park
MD, 20742
01+301-405-1000
tcalico@umd.edu

Chen Li
Educational Testing Service
Princeton
NJ, 08541
01+609-734-5993
cli@ets.org

ABSTRACT
We explore how data generated by an online formative automated writing evaluation tool can help connect student writing product and processes, and thereby provide evidence for improvement in student writing. Data for 12,337 8th grade students were retrieved from the *Criterion* database and analyzed using statistical methods. The data primarily consisted of automated holistic scores on the student writing samples, and the number of attempts on a writing assignment. The data revealed trends of positive association between the number of revisions and the mean writing scores. User logs were sparse to support study of additional behaviors related to the writing processes of planning and editing, and their relation to the writing scores. Implications for enhancing automated scoring based feedback with learner analytics based information are discussed.

Keywords
Automated scoring, learner analytics, formative writing, automated feedback, process and product

1. INTRODUCTION
The *Criterion*® *Online Writing Evaluation Service* [3], is a web-based writing tool that allows easy collection of writing samples, efficient scoring, and immediate feedback through the *e-rater*® automated essay scoring (AES) engine [2].

Criterion supports essay writing practice with a library of more than 400 essay assignments in multiple discourse modes (expository and persuasive) for students in elementary, middle, and high schools as well as in college. These prompts are used for classroom writing assignments and their scoring is supported by AES models. As a formative writing tool, *Criterion* has several features to facilitate writing processes and help learners improve their writing. These include planning templates, immediate feedback, multiple attempts to revise and edit, and resources such as a Writer's Handbook, a spell checker, a thesaurus and sample essays at different score points. The holistic scoring and feedback in *Criterion* is supported by *e-rater*. The analyses of errors and feedback are available for linguistic features of grammar, usage, mechanics, style and organization and development. There are limited studies on the pedagogical effectiveness of *Criterion* and AES systems in general [1, 5], and examining relation of product and process data for assessing writing quality [4]. Our motivation for this study was to analyze product data (holistic scores) in relation to process data (for revising) to provide evidence for effectiveness of the tool and automated feedback and scoring for improving writing. We report the observed trends for association between the two types of data, the cautions warranted in making strong claims based on these data, and the next steps.

2. METHODS
Data were extracted for 8th grade students for one school year from the *Criterion* database. The data spanned 295 days, and included 12,337 students from 183 schools; a total of 95,261 attempts were made across 41,473 assignments on 2,447 prompts.

Mean holistic scores by the *assignment* and by the *attempt* were examined to relate the revising behavior with improvement in writing scores. The results from the assignment and the attempt level analyses can easily be preliminary indicators of the tool's usefulness and effectiveness, and enhanced data logging capabilities of student actions in the system can provide richer information on writing processes.

3. RESULTS
3.1 Assignment Level
Of the 12,337 students who submitted assignments in the system, a little over 4,000 students submitted only one assignment over the full school year. About half of the students (N=6,663) completed a total of 2 to 6 assignments. A handful of students submitted as many as a total of 15 assignments. We identified groups of students who completed 2 to 5 unique assignments over the period of the full school year (the Ns were small for groups of students completing 6 or more assignments and hence excluded). The assignments in *Criterion* can be scored on a 4-point or a 6-point scale. We analyzed the data for responses evaluated on a 6-point scale only, and hence after filtering out the responses scored on the 4-point scale, the remaining sample size was 5,235 students. It should be noted that within each assignment, a user can have multiple attempts.

Figures 1a and 1b present the trends for the mean writing scores across assignments for the different groups based on the first attempt and the last attempt on the assignment, respectively. We draw quite a few interesting observations from the two graphs. The mean writing scores on the last attempt are always higher than the mean writing scores on the first attempt across all the assignments. Further, the mean writing score on the last attempt of the first assignment (first data point in Figure 1b) is almost always higher than the mean writing score on the first attempt of the fifth assignment (last data point in Figure 1a), suggesting that multiple attempts on an assignment is associated with a higher mean writing score than the total number of assignments completed by a user in the system.

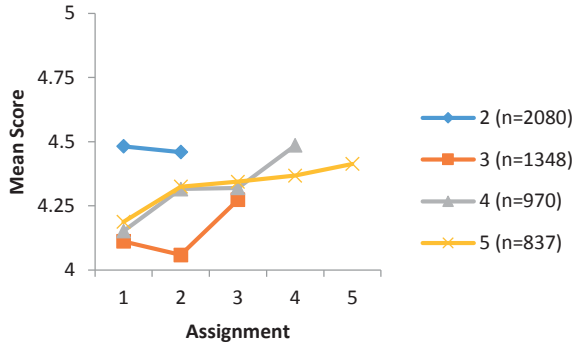

Figure 1a. Mean holistic score on the first attempt, per ordered assignment conditioned on total number of assignments

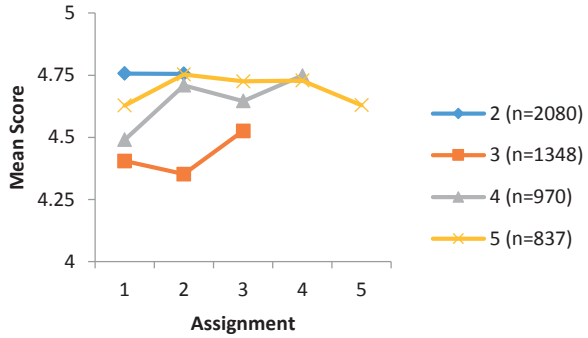

Figure 1b. Mean holistic score on the last attempt, per ordered assignment conditioned on total number of assignments

3.2 Attempt Level

After filtering for responses evaluated using 6-point scale, a total of 34,196 completed attempts were recorded in the system over the full school year. 15,841 of these attempts were instances of one attempt only per assignment. A few students completed as many as 10 attempts on an assignment which is the maximum limit by default. We identified groups of 2 to 6 attempts per assignment that included 16,846 instances (the Ns were small for groups of 7 or more attempts and hence excluded). Figure 2 presents the trends of mean writing scores across attempts for the different groups. The uniform trend of increase in the mean writing scores across the attempts for all the groups once again suggests that the revising process is associated with gains on the writing scores.

4. LIMITATIONS

The data on which trends have been reported were derived from a non-experimental setting. Large groups of students completed only one assignment or submitted only one attempt. Students who did engage in multiple assignments and/or multiple attempts hint at self-selection. The data are unbalanced and highly non-normal, and hence do not support rigorous statistical analyses but rather only lend themselves to exploration for trends.

Server log files were sparse for digital traces of student actions to support nuanced analyses of the corresponding writing processes. Information on students such as background variables is not available in the system. We analyzed data for only one grade level, but it would be of interest to examine if and how the trends based on product data as well as students' usage of the system vary across the different grade levels. Similar analyses of linguistic feature values or error analyses on the product can provide further insight into the process of improvement in student writing.

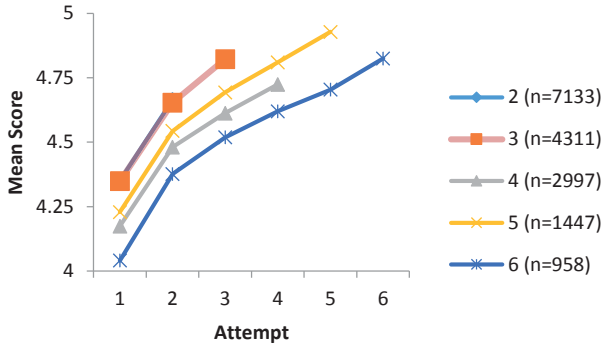

Figure 2. Mean holistic score, per ordered attempt by total number of attempts

5. CONCLUSION

Data currently available from *Criterion* are primarily on the work product; limited data are available for writing processes based on user actions. The additional data from our ongoing work on extension of *Criterion* to capture extended learner usage data will support further analysis of associations between the writing product and the processes, and their relation to change in student writing ability over time. This work has implications for extending application of automated scoring systems in formative contexts with the potential to provide richer feedback on product as well as processes, and enhancing the validity argument for automated scores as supported by response process data.

6. REFERENCES

[1] Attali, Y. 2004. Exploring the feedback and revision features of Criterion. Paper presented at the National Council on Measurement in Education, San Diego, CA.

[2] Attali, Y. & Burstein, J.C. 2006. Automated essay scoring with e-rater v.2. Journal of Technology, Learning, and Assessment, 4(3), (2006), 1–31.

[3] Burstein, J.C., Chodorow, M, & Leacock, C. 2004. Automated essay evaluation: the Criterion online writing service. AI Magazine 25(3), (2004), 27–36.

[4] Deane, P. 2014. Using writing product and process features to assess writing quality and explore how those features relate to other literacy tasks. ETS Research Report No. 14-03. Princeton, NJ: ETS.

[5] Foltz, P., Rosentsein, M., Dronen, N., & Dooley, S. 2014. Automated feedback in a large-scale implementation of a formative writing system: Implications for improving student writing. Paper presented at the annual meeting of the American Educational Research Association, Philadelphia, PA.