Listening. Learning. Leading.®

Research Report
ETS RR–15-26

Exploring the Feasibility of Using Writing Process Features to Assess Text Production Skills

Paul Deane

Mo Zhang

December 2015

Discover this journal online at
Wiley Online Library
wileyonlinelibrary.com

ETS Research Report Series

EIGNOR EXECUTIVE EDITOR

James Carlson
Principal Psychometrician

ASSOCIATE EDITORS

Beata Beigman Klebanov
Senior Research Scientist - NLP

Heather Buzick
Research Scientist

Brent Bridgeman
Distinguished Presidential Appointee

Keelan Evanini
Senior Research Scientist - NLP

Marna Golub-Smith
Principal Psychometrician

Shelby Haberman
Distinguished Presidential Appointee

Donald Powers
Managing Principal Research Scientist

Gautam Puhan
Principal Psychometrician

John Sabatini
Managing Principal Research Scientist

Matthias von Davier
Senior Research Director

Rebecca Zwick
Distinguished Presidential Appointee

PRODUCTION EDITORS

Kim Fryer
Manager, Editing Services

Ayleen Stellhorn
Editor

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Exploring the Feasibility of Using Writing Process Features to Assess Text Production Skills

Paul Deane & Mo Zhang

Educational Testing Service, Princeton, NJ

In this report, we examine the feasibility of characterizing writing performance using process features derived from a keystroke log. Using data derived from a set of *CBAL*™ writing assessments, we examine the following research questions: (a) How stable are the keystroke timing and process features across testing occasions? (b) How consistent are the patterns of feature–human correlation across genres and topics? (c) How accurately can we predict human ratings on writing fundamentals using a combination of the keystroke timing and process features, and what are the contributions of each feature to the reliable variance in the human ratings? (d) If we train a predictive model on one prompt, how well do its predictions generalize to the other prompts of the same or different genre? The results of the study indicate that keystroke log features vary considerably in stability across testing occasions and display somewhat different patterns of feature–human correlation across genres and topics. However, using the most stable features, we can obtain moderate to strong prediction of human essay scores, and those models generalize reasonably well across prompts though more strongly within than across writing genres.

Keywords writing; writing process; writing assessment; keystrokes; keystroke log; process features; writing quality

doi:10.1002/ets2.12071

Writing involves multiple, coordinated cognitive processes (Alamargot & Chanquoy, 2001; Alamargot & Fayol, 2009; Bereiter & Scardamalia, 1987; Hayes & Flower, 1980). The literature recognizes a number of distinct writing subprocesses, such as translating ideas into words, transcribing words onto the page, and evaluating the quality of text produced to date (Chenoweth & Hayes, 2003). The literature on the role of timing in the writing process (Chenoweth & Hayes, 2003; Hayes & Flower, 1980; Perl, 1979; Torrance & Galbraith, 2006) indicates that pause patterns reflect specific subprocesses such as planning and evaluation that interrupt fluent text production. Unskilled writers may be restricted by limitations in a subprocess, such as transcription fluency, which may limit capacity to learn/exercise other writing skills (Berninger, 1996; Flower & Hayes, 1981; Kellogg, 2001, 2008; McCutchen, 1996, 2000). Thus, the distribution and patterning of pauses (or, conversely, of bursts of fluent text production) provides indirect evidence about underlying cognitive processes.

Several broad trends have been observed relating pause patterns during writing with the quality of the resulting text (Alves, Branco, Castro, & Olive, 2012; Connelly, Dockrell, Walter, & Critten, 2012; Hayes, 2012; Kaufer, Hayes, & Flower, 1986; Miller, 2000; van den Bergh & Rijlaarsdam, 2001; Wengelin, 2006). Similar results have been obtained in prior published research at Educational Testing Service (ETS), including studies by Almond, Deane, Quinlan, Wagner, and Sydorenko (2012), Deane and Quinlan (2010), and Deane, Quinlan, and Kostin (2011). In particular, for stronger writers, text tends to be produced efficiently in longer bursts; pauses are more likely to happen at natural loci for planning such as clause and sentence boundaries, and more editing and revision behavior can be observed. On the other hand, for weaker writers, text tends to be produced less efficiently, and pauses appear in locations that suggest difficulties in typing, spelling, word-finding, and other transcription processes.

However, these relationships are complicated by a variety of factors, including (a) developmental shifts (for instance, fluency of transcription increases with age; Abbott, Berninger, & Fayol, 2010; Berninger, 1999; Graham, Berninger, Abbott, Abbott, & Whitaker, 1997; Limpo & Alves, 2013); (b) differences between tasks or writing genres (for instance, writing processes for narrative, argument, and exposition may differ; Beauvais, Olive, & Passerault, 2011); (c) differences in temporal distributions within a task (for instance, the value of planning may be higher earlier in a writing session and be counterproductive toward the end of a session; Breetvelt, van den Bergh, & Rijlaarsdam, 1994); and (d) differences in

Corresponding author: P. Deane, E-mail: pdeane@ets.org

personal style (for instance, some people may prefer extensive advance planning whereas others may prefer to free write and revise; Tillema, van den Bergh, Rijlaarsdam, & Sanders, 2011).

A number of programs for collecting keystroke log data have been developed for the purpose of measuring writing patterns. These include Inputlog, developed for use in studying multimodal professional writing environments (Leijten & van Waes, 2006, 2013; van Waes & Leijten, 2006); Scriptlog, developed for use in experimental psycholinguistic research (Andersson et al., 2006; Strömqvist, Holmqvist, Johansson, Karlsson, & Wengelin, 2006); and Translog, developed for use in translation research (Jakobsen, 2006). ETS has developed its own keystroke logging program (Almond et al., 2012). All of these programs rely on essentially similar methods, which involve capturing character input and tracking the length of time between inputs. Several families of methods for analyzing keystroke logs are well attested in the literature, including both categorical and statistical methods. Categorical methods include pause analysis (analyzing the distribution of pauses above a minimum length), revision analysis (analyzing what chunks are inserted or deleted), burst analysis (classifying bursts into different types such as those occurring before/after revision), and possibly expanding the log with some information about the linguistic state of the text (Leijten, Hoste, Van Waes, Macken, & Van Horenbeeck, 2012). Statistical methods include calculation of summary features, such as burst length and pause duration, with some work using mixture models and principal component analysis (Almond et al., 2012; Baaijen, Galbraith, & de Glopper, 2012; Deane, 2014).

Deane and Quinlan (2010), and Deane, Quinlan, & Kostin (2011) explored ways in which writing timing and process features can function as predictors of writing scores, but Deane (2014) found that issues of stability and generalization might exist, at least for some keystroke features. In particular, Deane (2014) examined the relationship between keystroke features from essays instantiating very different genres and topics and found relatively low correlations of keystroke features across paired student essays addressing different topics and genres combined with somewhat different patterns of correlations with external variables.

In an exploratory factor analysis, Deane (2014) identified three factors supported by keystroke features: measuring latency (or hesitation in text production), editing behaviors, and burst span (amount of text produced without need for a longer pause). Within and between prompts, all three factors showed similar patterns of correlation with writing score (a significant negative correlation for latency and significant positive correlations for editing behaviors and burst span). However, the size of the effects varied noticeably. For example, the correlation of the editing factor with human scores was .32 in one prompt but only .18 in the other; conversely, in one prompt, the correlation of the burst span factor with human scores was .09 in one prompt and .23 in the other. Moreover, the relation to external reading measures varied systematically. In one prompt the latency factor was significantly and negatively correlated with a variety of reading measures administered to the same students, and the editing and burst span factors were not statistically significant. By contrast, in the other writing prompt, the editing and burst span factors had significant positive correlations with the same reading measures, but the latency factor was not significantly correlated. Deane (2014) thus suggested that while evidence can be extracted from keystroke logs about characteristics of student performance, there is enough variability in performance to raise questions about the stability and generalizability of the patterns identified.

The current study examines a large pool of essays, following a design intended to sample multiple topics within two different genres. It therefore provides an important opportunity to examine the extent to which features extracted from keystroke logs generalize across prompts and provide reliable indicators of fundamental writing skill as measured by human raters. It is intended as an exploratory study, focused on the following four research questions.

- Research Question 1 (RQ1): How stable are the keystroke timing and process features across testing occasions?
- Research Question 2 (RQ2): How consistent are the patterns of feature – human correlation across genres and topics? That is, how well can models based on keystroke features predict performance on the fundamental text production rubric?
- Research Question 3 (RQ3): How accurately can we predict human ratings on writing fundamentals using a combination of the keystroke timing and process features, and what are the contributions of each feature to the reliable variance in the human ratings?
- Research Question 4 (RQ4): If we train a predictive model on one prompt, how well do its predictions generalize to the other prompts of the same or different genre?

This study purposefully did not examine the relationship between process/keystroke timing features and the product features used in automated essay scoring systems. These issues have been explored elsewhere (e.g., Deane, 2014;

Deane, Quinlan, & Kostin, 2011), and we expect to explore them in greater depth in future publications (e.g., Zhang & Deane, in press). Our immediate goal in this study was to examine the extent to which writing process features provide measurement of fundamental text production skills, without reference to features of the final written product.

Method

Instrument

Six English-language arts (ELA) test forms were developed as part of the *CBAL*™ learning and assessment tool research initiative at ETS (Bennett, 2011; Bennett & Gitomer, 2009; Deane, Fowles, Baldwin, & Persky, 2011). As these publications discuss, the CBAL initiative is designed to explore ways to build high-quality assessments that are learning experiences in their own right and can also provide useful information to guide instruction. For present purposes, the following salient characteristics are worth noting.

- Each CBAL ELA test form started with a preliminary, or lead-in, section that required students to read, think, and respond to questions about a set of source documents.
- After the lead-in section, students were required to complete an essay task in which they responded to the same source documents. The two parts of a test form were linked by a common purpose, which also defined the genre of the essay that students were required to write.
- Each part was intended to be administered in a single 45-minute class session for a total of 90 minutes for a test form.

The six test forms were designed to sample two different purposes for writing (or genres). One writing purpose focused on building an argument using evidence extracted from secondary sources, and the second focused on evaluating two proposals using explicit evaluation criteria and recommending one proposal over another on that basis. While both genres focused broadly on building effective arguments, the specific genres required the writer to achieve different rhetorical goals that entail distinct organizational patterns. Within each genre, we sampled three topics, as shown in Table 1.

Participants

Six writing test forms were administered during the spring of 2013 to a convenience sample of more than 2,500 students from Grade 6 to Grade 9 in seven US states. The students who took the forms focused on writing an argument essay were drawn primarily from an eighth-grade cohort (1,129 students) with a smaller sample drawn from adjacent grades (404 from seventh grade and 708 from ninth grade). The students who took the form focused on writing a recommendation were drawn primarily from a seventh-grade cohort (930 students) with a smaller sample drawn from adjacent grades

Table 1 Description of the Test Forms

Unifying purpose of assessment	Genre (of the final essay task)	Topic	Test form (acronym)
Oppose or support a policy recommendation	Argument essay	Should the United States ban advertising to children under 12?	Ban ads (BA)
		Should schools pay students for getting good grades?	Cash for grades (CG)
		Should schools recommend that parents restrict students' use of social networking?	Social networking (SN)
Decide which of two alternative plans best satisfies explicit decision criteria	Written recommendation	What would be the best choice of service learning project for a class to carry out?	Service learning (SL)
		What would be the best theme for a school culture fair?	Culture fair (CF)
		What is the best way for a school to spend a large sum of money provided by a generous donor?	Generous gift (GG)

(137 from sixth grade and 226 from eighth grade). In both cases, slightly over half of the students (54% and 56%) were drawn from school districts in Idaho, where ETS had an opportunity for a large-scale data collection, supplemented by data from additional schools recruited from other states. A smaller sample was collected for two forms (ban ads [BA] and service learning [SL]) than for the remaining forms, because those two forms had already been piloted in prior studies (e.g., Deane, 2014; Fu, Chung, & Wise, 2009).

In the argument essay administration, demographic data were available for 88% of students, of whom 51.3% were female and 48.7% were male; 71.8% were White, 19.7% were Hispanic, 3.3% were African American, 3.5% were Asian, and less than 1% belonged to any other group. Further demographic data were available for 1,888 students, of whom 97.8% were initially English proficient, 1.6% were English-language learners (ELLs), and less than 1% percent were reclassified. Annual yearly progress (AYP) data were available for 1,204 students, of whom 3.7% required AYP accommodation. Socioeconomic status (SES) data were available for 1,042 students, of whom 41.6% qualified for free or reduced lunch.

In the written recommendation administration, demographic data were available for 81% of students, of whom 49.4% were female and 50.6% were male; 68.4% were White, 22.9% were Hispanic, 4.6% were African American, 3.5% were Asian, and less than 1% belonged to any other group. Further demographic information was available for 1,144 students; of these students, 92.7% were initially English proficient, 4.3% were reclassified English proficient, and 2.9% were ELL. AYP data was available for 1,203 students, of whom 3.6% required accommodation and 40.7% qualified for free or reduced school lunch programs.

Test Administration

Parental permission was obtained in advance, and schools were compensated for each completed test form. Each test form required two class sessions (for the lead-in and essay tasks, respectively). The two sessions were scheduled within 1 week of one another, usually consecutively on the same day. Students answered all questions, including writing their essays, in an electronic form using an online interface. It was expected and stated in the test administration manual that tests were given in regular classes, and students were not expected to leave the test until completion. Keystroke logs were collected as part of data collection.

Each student took two of the three test forms focused on the same genre, with the second form being administered within 2 weeks after the completion of the first form. Students were randomly assigned within classes to one of the three pairings of the test forms (Table 2) within a genre.[1] An earlier study (Deane, 2014) examined patterns of performance across very distinct genres; in this study, by contrast, we focused on the effect of changes in topic within each genre. As each form required two separate class sessions to complete, not all students wrote both assigned essays. If a student missed either of the two writing sessions, or produced a 0-coded response, that student's data were excluded from Table 2, which shows the number of students who completed both essays for each essay pair. Of the 575 students, 439 completed both BA and another essay; 667 of 900 completed the cash for grades (CG) essay and another essay; 618 of 805 completed the social network (SN) essay and another essay; 562 of 659 completed SL and another essay; 723 of 832 completed the culture fair (CF) and another essay; and 710 of 825 completed the generous gift (GG) essay and another essay. Due to the sampling scheme, we obtained much larger samples completing the CG – SN and CF – GG pairs, which should be considered when interpreting subsequent analyses.

Table 2 Number of Essays in Each Essay Pairing

Genre	Form pair	*n* of paired essays
Argument essay	BA – CG	244
	BA – SN	195
	CG – SN	423
Written recommendation	SL – CF	249
	SL – GG	236
	CF – GG	474

Note: BA = ban ads; CG = cash for grades; SN = social networking; SL = service learning; CF = culture fair; GG = generous gift.

Scoring and Rater Characteristics

For the analyses in the current study, we are concerned only with the test session that administered the essay task. The essays were scored on two rubrics: one focused on writing fundamentals (usage and mechanics, language, organization, and development on a 1- to 5-point scale), and the other focused on the quality of the thinking (also on a 1- to 5-point scale). In both rubrics, a human score of 0 was used to denote essays with unusual response characteristics such as empty, off topic, plagiarism, or random keystrokes. Those responses were considered as outliers and were excluded from our analyses. Four of the test forms (CG, SN, CF, and GG) were scored twice on each rubric. Two of the test forms (i.e., BA and SL) had been administered and analyzed in prior studies, so only about 20% of the responses were double scored.

The raters employed in this study were teachers trained by ETS staff and not professional ETS scorers. Most of them were scoring the CBAL assessments for the first time. On the two forms with approximately 20% double scoring (i.e., BA and SL), we used professional ETS scorers for the second rating.

In this study, as the keystroke timing and process features are most construct relevant with respect to fundamental text production skills, we focused on their relation to human ratings on the writing fundamentals rubric. Our reasoning was as follows: The correlation between adjudicated scores on the two rubrics is .77 in the GG data set, .72 in the CF data set, and .68 for the 20% of the SL data set that was double scored. Despite the moderately strong correlations, keystroke features also appear to contribute to prediction of the argument quality score. For example, in the GG data set, if we attempt to predict performance on the content quality rubric and enter the writing fundamentals score first, the writing fundamentals score has a statistically significant beta weight of about 0.67, but three other variables discussed elsewhere in this paper also contribute to prediction: variability of sentence pauses (with a statistically significant beta weight of about 0.12), variability of burst lengths (with a statistically significant beta weight of about 0.11), and variability in burst times (with a statistically significant beta weight of about −0.08) for a change in R^2 of +.02. This result indicates that students who received high scores for the quality of content tended to have more long pauses between sentences and larger numbers of very long bursts, suggesting that these students were generally more fluent than the ones who received lower quality scores. This outcome is consistent with results in the literature, that is, McCutchen's (1996) capacity hypothesis, in which fluency at fundamental text production frees resources for other tasks such as critical thinking. But note that this account assumes that greater fluency is causally linked to students achieving higher content scores, so that the features remain, essentially, measures of fluency. We therefore focus in the rest of the paper on validating the use of these features as measures of an underlying skill (competency at fundamental text production), as measured by the writing fundamentals rubric.

Preliminary Analysis of Rater Performance

Table 3 shows the number of essays that were scored by a first and a second rater using the writing fundamentals rubric. Also shown in Table 3 are the means and standard deviations of the human scores for each test form and the interhuman agreements, expressed as Pearson correlation coefficient, quadratically weighed kappa, and exact and 1-point adjacent percentage agreement. The mean scores were slightly higher for the argument essay forms (ranging from 2.61 to 2.79) than for the written recommendation form (ranging from 2.40 to 2.65). The correlations range from a low of .48 to a high of .77 below that desirable for operational scoring. Given the exploratory nature of this study, we

Table 3 Within-Form Interhuman Agreement (Writing Fundamentals Rubric)

Genre	Form	*n* (H1, H2)	Mean (H1)	*SD* (H1)	Mean (H2)	*SD* (H2)	*r*	QWK	Exact %	Adj. %
Argument essay	BA	575,113	2.72	1.07	2.86	1.16	.55	0.55	34	85
	CG	900,900	2.79	1.00	2.77	0.99	.67	0.67	51	95
	SN	798,798	2.61	0.94	2.54	0.93	.48	0.48	42	91
Written recommendation	SL	562,108	2.50	1.00	2.92	0.92	.77	0.76	60	97
	CF	825,825	2.40	0.98	2.72	1.00	.67	0.64	47	94
	GG	823,823	2.65	0.97	2.37	0.85	.69	0.66	52	96

Note: All *r* (Pearson correlation coefficient) and QWK (quadratically weighted kappa) values were statistically significant at $p < .001$ level. Exact % = exact percentage agreement; adj. % = 1-point adjacent percentage agreement; *SD* = standard deviation; H1 = first human ratings on the writing fundamentals rubric; H2 = second human ratings on the writing fundamentals rubric. BA = ban ads; CG = cash for grades; SN = social networking; SL = service earning; CF = culture fair; GG = generous gift.

Table 4 Cross-Form Interhuman Correlation Coefficients

Genre	Form pair	H1/H1 (*n*)	H1/H2 (*n*)	H2/H2 (*n*)	H2/H1 (*n*)
Argument essay	BA – CG	.35 (244)	.30 (244)	.52 (46)	.24 (46)
	BA – SN	.30 (195)	.31 (195)	.52 (40)	.51 (40)
	CG – SN	.45 (423)	.40 (423)	.43 (418)	.41 (418)
Written recommendation	SL – CF	.38 (249)	.47 (249)	.49 (45)	.47 (45)
	SL – GG	.56 (236)	.54 (236)	.49 (50)	.62 (50)
	CF – GG	.43 (472)	.47 (472)	.40 (472)	.45 (472)

Note: Only about 20% of BA (ban ads) and SL (service learning) essays were double scored, resulting in smaller sample sizes for comparisons involving either prompt. The first value (.35) in the H1/H1 (first human rating) column represents the Pearson correlation coefficient between H1 on the writing fundamentals rubric in the BA prompt and the first human ratings (H1) on the CG prompt. CG = cash for grades; SN = social networking; CF = cultural fair; GG = generous gift.

considered the human rating quality acceptable for the purpose of evaluating the performance of keystroke timing and process features.[2]

Given the level of reliability that these correlations indicate, we would expect that the correlations between human ratings across forms would be somewhat lower but still fall generally in the moderate range because they would be affected by rater error on both prompts. As Table 4 indicates, these expectations are accurate. It appears that the human ratings provide an acceptable criterion variable to which keystroke log timing features can be related.

Data Cleaning

The timing and process features extracted from the keystroke logs are only likely to be statistically meaningful for essays that contain a significant number of keystroke events. Very short logs contain very little information (and are unlikely to be valid responses to an essay prompt in any case). A similar argument can be made for an essay that is composed within an extremely short amount of time. As a result, we eliminated essays with fewer than 25 words, if those had not already been assigned a score of 0 during human scoring. We therefore also eliminated a small number of outlier essays where the total time spent writing an essay was only a few seconds or where the ratio of text produced to time on task was very large (in excess of one word per second), suggesting that the entire essay had been pasted in with very little time spent on composition. In a small number of keystroke logs, it appeared that the writer had closed and reopened the browser while taking the test, resulting in a partial loss of data. These logs were also excluded from data analysis. The total number of essays eliminated was very small ($n = 32$ for BA, $n = 40$ for CG, $n = 48$ for SN, $n = 4$ for SL, $n = 8$ for GG, $n = 8$ for CF). All results reported in this paper (including the data in Tables 2 – 4) are based on the final cleaned data sets.

Keystroke Timing and Process Feature Extraction

The features characterizing writing process based on the keystroke logs were automatically extracted using the methods documented in Almond et al. (2012). In this approach, the keystroke log is used to identify behavioral features (e.g., bursts of text production vs. long pauses; various editing events, including backspacing, insertions, and deletions) and textual boundaries (e.g., between words, sentences, and paragraphs). The following list indicates the primary features identified:

- Total time on task (milliseconds)
- Length of bursts (based on sequences with no pause > 2/3 seconds; in words)
- Duration of bursts (in milliseconds, logged)
- Duration of pauses between paragraphs (in milliseconds, logged)
- Duration of pauses between sentences (in milliseconds, logged)
- Duration of pauses between words (in milliseconds, logged)
- Duration of pauses between characters within a word (in milliseconds, logged)
- Duration of cut/paste/jump events (in milliseconds, logged)
- Duration of multiple backspace events (in milliseconds, logged)
- Duration of pauses before a single-character backspace (in milliseconds, logged).

Based on the literature on pause patterns in writing, we offer the following interpretations:

- Total time on task corresponds to the overall level of effort put into producing the text.
- Greater variability of between-sentence pauses and in the distribution of cut, paste, and jump events can represent the cognitive states of deliberation, planning, and editing.
- Greater variability of between-word and within-word pauses and backspacing may reflect difficulties in word-finding, spelling, or typing.
- Burst duration does not have as obvious an interpretation, but greater variability in the time spent in bursts of text production could imply increased fluency that has an impact on the availability of cognitive resources for higher level processing.

For each keystroke timing and process feature, we calculated three summary values: the mean, the standard deviation of durations, and the normalized amount of the total time that each event type occupies in the keystroke timing log. However, preliminary analyses indicated that one version of the summary feature — the standard deviation of pause durations, rather than the average or normalized value — generally showed more consistency in size and magnitude across tasks for the same student, possibly because of the highly skewed nature of the underlying distributions (Almond et al., 2012; Deane, 2014). In the analyses presented hereafter, we therefore focused on this class of summary feature.

Data Analyses

Data Analyses for Research Question 1

RQ1 states as follows: How stable are the keystroke timing and process features across testing occasions?

To answer RQ1, we computed the Pearson product – moment correlations between the same features for the same individuals on different test forms. We then inspected the correlation patterns to identify differences in feature performance between the argument essay and written recommendation genres.

If a feature is reliable, we would expect some stability in its values across testing occasions. A high correlation would mean that the pattern of feature values (though not perhaps the absolute magnitudes) reflects a stable characteristic of the writing process of an individual. The lower the correlations are, the more the value of the feature may be determined by specific properties of individual forms or testing occasions. Moreover, if a feature shows strong stability across test forms in one genre, but little generalization in another, that feature may be reflecting differences in writing patterns that reflect task differences rather than variations in individual ability.

Data Analyses for Research Question 2

RQ2 is stated as follows: How consistent are the patterns of feature – human correlation across genres and topics? That is, how well can models based on keystroke features predict performance on the fundamental text production rubric?

To provide initial descriptive analyses, we computed the Pearson product – moment correlations of each of the keystroke timing and process features with human scores on writing fundamentals rubric and compared the patterns extracted across test forms and genres. The keystroke timing and process features reflect part — but only part — of the writing fundamentals rubric. As such, we would expect weak to moderate, and not strong, correlations between individual features and human ratings. However, the usefulness of keystroke timing and process features as measures of an underlying trait of fundamental text production skill depends upon the consistency and magnitude of feature correlations across raters and forms. If the correlations are significant and in the same direction on different forms, then we can reasonably hypothesize that they reflect an underlying trait associated with writing expertise. To the extent that the patterns are consistent across genres, we would have evidence that supports some measurement of a common underlying trait. On the other hand, if the patterns of correlations are consistent within one genre but very different from those observed in the other genre, we may be able to conclude that the features are being affected by task requirements.

Data Analyses for Research Question 3

RQ3 is stated as follows: How accurately can we predict human ratings on writing fundamentals using a combination of the keystroke timing and process features, and what are the contributions of each feature to the reliable variance in the human ratings?

For each prompt, we randomly divided the sample into model-building and evaluation data sets. Using the model-building sample, we regressed the first human ratings on the keystroke timing and process features using stepwise regression and retaining only significant features. The stepwise feature selection we used was a modification of the forward-selection method in that the features entered into the model did not necessarily stay in the model. The stepwise selection method examined all the features already included in the model and eliminated any feature that did not produce a significant *F*-measure statistic. Only after this step and the necessary eliminations were completed can another feature be added into the model. In the end, all the features resulting in the final model accomplished a significant *F*-measure statistics at the $p < .01$ level, and every feature in the model was significant at the $p < .05$ level.

We used the *R*-squared value resulting from the models as an indication of the collective predictive value of the keystroke timing and process features with respect to human ratings on writing fundamentals (i.e., fundamental text production skills). To the extent that the models we built to account for a considerable amount of the total variance, we could reasonably conclude that the patterns of keystroke features are giving us insight into writers' fundamental writing skills.

We then examined the contribution of each keystroke timing and process feature to the reliable variance in explaining the human ratings. Such analyses would provide us further understanding on the stability of the features within and across the two genres.

Data Analyses for Research Question 4

RQ4 is stated as follows: If we train a predictive model on one prompt, how well do its predictions generalize to the other prompts of the same or different genre?

We applied those prompt-specific models to the prompt on which the model was constructed and to the other prompts from the same and different writing genres. Using independent samples, we then evaluated and compared the models' performance on the base prompt with its performance on the other prompts using the Pearson correlation coefficient between resulting predicted scores and the human ratings (on the writing fundamentals rubric). If the keystroke feature model weights generalize, we would expect the correlations of the resulting predicted scores with human ratings remain as strong (or almost as strong) as when applied to other test forms. Further, if they are stronger for other forms in the same genre than they are for forms in the other genre, that situation would indicate that the keystroke features generalize better within genre than across genre with regard to predicting human scores.

We also conducted a less direct comparison, in which we calculated Pearson correlation coefficients of predicted scores on one form with human scores for the same student on a different form. If the keystroke feature model weights generalize well, we would expect the predicted scores (derived from models based on those features) on one test form to correlate at least moderately with the same student's score on a parallel form, without large differences in the correlation between the automated model and human scores on the same versus different prompts. Or, to put the point more precisely, we would expect the predicted scores on Prompt A resulting from the model based on Prompt A to correlate with the human ratings for the same students on Prompt B in a similar way as the predicted scores on Prompt B (resulting from the model based on Prompt B) correlate with human ratings of the same students' performance on Prompt A. A lack of comparability may suggest a lack of generalization.

Results

Results for Research Question 1

Correlations Between Features Across Test Forms

RQ1: How stable are the keystroke timing and process features across testing occasions? The results for the patterns of correlations between the same features for the same individuals on different test forms are given in Table 5.

Total Time on Task

The amount of time during which an individual was actively engaged in the writing task was moderately positively correlated across test forms. Correlations between the total time on task for the same individual on two different forms

Table 5 Cross-Form Correlations of Keystroke Timing and Process Features

	Argumentative essay form pairs				Written recommendation form pairs			
Timing and process feature	BA – CG	BA – SN	CG – SN	Average *r*	SL – CF	SL – GG	CF – GG	Average *r*
Total time on task	.52*	.58*	.61*	.57	.49*	.47*	.49*	.48
SD (burst length in words)	.62*	.62*	.48*	.57	.85*	.76*	.80*	.80
SD (burst duration in milliseconds)	.35*	.15*	.29*	.26	.43*	.57*	.37*	.46
SD (pause between sentences)	.19*	.22*	.24*	.21	.09	.12	.22*	.14
SD (pause within words)	.62*	.48*	.39*	.50	.74*	.60*	.67*	.67
SD (pause between words)	.49*	.30*	.34*	.38	.13*	.43*	.44*	.33
SD (duration of cut/paste/jump events)	.09	.09*	.26*	.15	.20*	.22*	.20*	.21
SD (duration of single backspace event)	.18*	.10*	.20*	.16	.17*	.23*	.17*	.19
SD (duration of multiple backspace event)	.08	.01	.18*	.09	.22*	.22*	.18*	.21

Note. The results are based on the first human ratings on the writing fundamentals rubric in each test form. The sample size for each form pair can be found in Table 2. Average *r* = unweighted average of the three values within a genre. BA = ban ads; CG = cash for grades; SN = social networking; SL = service learning; CF = cultural fair; GG = generous gift.
* indicates significance at $p < .05$ level.

were .52, .58, and .61 with an average of .57 for the argument essay form set, but they were slightly lower for the written recommendation form set, with the values being .49, .47, and .49 for the three pairs with an average of .48.

Burst Length

Burst length (as measured by the standard deviation of the number of words produced without a long pause) patterned somewhat similarly to the total time on task for two of the three form pairs in the argument essay form set and exhibited considerably lower interform correlations for one form pair in the argument essay form set (i.e., CG – SN). For all three form pairs in the written recommendation form set, the cross-form correlations for this feature were considerably higher than the correlations for total time on task. Overall, the standard deviations of burst length in words were

- moderately and positively correlated between pairs of forms within individuals for the argument essay form set; and
- highly correlated between pairs of forms for the written recommendation form set.

The average correlations were .57 for the argument essay genre and .80 for the written recommendation genre.

Burst Duration

Burst length was also measured in terms of time duration in milliseconds. This feature, though, appeared to be less stable than burst length. For all pairs of test forms in both genres, the correlations were noticeably lower for burst length in milliseconds than by number of words. The average correlation was only .26 for the argument essay form set, with the lowest being .15 for the BA – SN pair. The average correlation for the other genre, written recommendation, was .46, also considerably lower than the value obtained from the other burst length-related measure. However, relatively speaking, this feature appears to be more stable for the written recommendation form set than for the argument essay form set. The lowest correlation for a pair of written recommendation forms (.37 for CF – GG) is higher than the highest correlation for a pair of argument essay forms (.35 for BA – CG).

Standard Deviations of Keystroke Timing Events Normalized Against Total Time on Task

These features varied considerably in their level of stability within genre. For example, in the argumentative essay form set, the average correlation for pauses within words is .50. In contrast, cut/paste/jump events had only an average correlation of .15. The same phenomenon was found for the other genre, where the correlation was .67 for pauses within words but much lower for cut/paste/jump events. In the argument essay form set, the average correlational values across all these features ranged from as low as .09 for multiple backspace events to as high as .50 for pauses within words. On the other hand, the average value of these correlations in the written recommendation form set ranged from .14 for pauses between sentences to .67 for pauses within words. Further, these features showed the greatest differentiation between genres. Four

Table 6 Within-Form Correlations Between Keystroke Timing and Process Features and Human Ratings on Writing Fundamentals

	Argument essay forms						Written recommendation forms					
	BA		CG		SN		SL		CF		GG	
Measure	H1	H2	H1	H2	H1	H2	H1	H2	H1	H2	H1	H2
Total time on tasks	.50*	.49*	.52*	.51*	.48*	.44*	.44*	.54*	.48*	.45*	.52*	.50*
SD (burst length in words)	.20*	.28*	.15*	.17*	.13*	.13*	.28*	.23	.20*	.25*	.27*	.29*
SD (burst duration in milliseconds)	.04	**−.05**	.05	.05	.00	.00	**−.24***	**−.13**	**−.19***	**−.18***	**−.14***	**−.13***
SD (pause between sentence)	.21*	.15	.21*	.22*	.28*	.21*	.20*	.36*	.19*	.23*	.25*	.25*
SD (pause within words)	**−.05**	**−.08**	**−.09***	**−.11**	**−.07**	**−.06**	**−.22***	**−.02**	**−.14***	**−.20***	**−.17***	**−.16***
SD (pause between words)	.07	**−.02**	.06	.02	.02	.04	**−.12***	.02	**−.01**	**−.04**	**−.02**	**−.03**
SD (duration of cut/paste/jump events)	.23*	.18	.30*	.32*	.27*	.25*	.22*	.30*	.23*	.23*	.31*	.27*
SD (duration of single backspace event)	.01	**−.14**	.15*	.15*	.03	.03	**−.11***	.02	**−.06**	**−.03**	.05	.04
SD (duration of multiple backspace event)	.14*	.08	.12*	.12*	.14*	.11*	.03	.18	.06	.08*	.13*	.11*

Note. Values in bold contain negative values. The sample size for each form can be found in Table 3. H1 = first human rating on the writing fundamentals rubric; H2 = second human rating on the writing fundamentals rubric; BA = ban ads; CG = cash for grades; SN = social networking; SL = service learning; CF = cultural fair; GG = generous gift.

* indicates significance at $p < .05$ level.

of the six features had a higher average correlation in the written recommendation form set than in the argument essay form set, and in some cases, the discrepancy was rather noticeable (e.g., .50 vs. .67 for pauses within words).

Results for Research Question 2

RQ2: How consistent are the patterns of feature–human correlation across genres and topics? That is, how well can models based on keystroke features predict performance on the fundamental text production rubric?

We examined the consistency of the feature–human correlation across genres and writing topics. Table 6 presents the Pearson correlation coefficients of each keystroke timing and process feature with the human scores on writing fundamentals.

Note that negative correlations are to be expected for certain features on construct grounds. Behaviors that reflect difficulties in word-level text production, such as word-finding and spelling-correction behaviors, may lead to greater variability in the duration of within-word (and possibly, between-word) pauses and single-character backspacing events. The negative correlations for standard deviation of burst durations are less clearly interpretable. On the other hand, there were fairly consistent and significant positive correlations for total time on task, standard deviation of burst length, standard deviation of between-sentence pauses, and standard deviation of cut/paste/jump events, all of which can be interpreted as reflecting greater overall fluency and the presence of editing behaviors. The low frequency of certain events (e.g., single and multiple backspacing) might have attributed to the low feature–human correlations for some features.

Notably, the burst duration feature showed contrasting performance in the two genres. Correlations of standard deviation of burst duration with human score were not statistically significant (and near zero) for argument essay prompts, but were statistically significant (and negative) for written recommendation prompts.

Results for Research Question 3

RQ3: How accurately can we predict human ratings on writing fundamentals using a combination of the keystroke timing and process features, and what are the contributions of each feature to the reliable variance in the human ratings?

The overall performance of keystroke features varies somewhat by genre. Models built for the three argument essay–focused assessments achieved correlations with human scores between .55 and .61. By contrast, the models built for the three test forms focused on policy recommendations achieved correlations with human scores between .66 and .68.

Table 7 provides an estimate of how much variance in fundamental text production skill (as measured by first human ratings) can be accounted for by each keystroke timing/process feature. Also provided in the table are the regression statistics associated with final models based on each model-building data set.

Table 7 Final Stepwise Regression Model Statistics and Relative Feature Weights by Prompt

		Argument essay forms			Written recommendation forms		
		BA	CG	SN	SL	CF	GG
Model statistics	*n*	288	450	403	281	416	413
	R-squared	.36	.39	.36	.45	.39	.48
	Adj. *R*-squared	.35	.38	.35	.44	.38	.47
	F-measure	31.24	47.32	37.65	44.83	37.47	53.06
	p value	<.001	<.001	<.001	<.001	<.001	<.001
	Root MSE	0.87	0.79	0.78	0.74	0.74	0.72
Feature weights	Total time on tasks	48.3%	41.0%	43.9%	38.6%	36.8%	38.5%
	SD (burst length in words)	22.4%	19.1%	22.1%	24.1%	21.1%	19.7%
	SD (burst duration in milliseconds)	7.2%	10.5%	6.5%	20.0%	17.5%	18.1%
	SD (pause between sentence)	—	—	—	—	—	5.1%
	SD (pause within words)	—	12.6%	13.3%	10.8%	5.5%	6.9%
	SD (pause between words)	11.8%	—	—	—	—	—
	SD (duration of cut/paste/jump events)	—	10.2%	5.5%	—	6.9%	7.7%
	SD (duration of single backspace event)	10.3%	6.5%	8.8%	6.5%	5.3%	4.0%
	SD (duration of multiple backspace event)	—	—	—	—	6.9%	—

Note. — indicates that particular feature did not enter the final model. Results in this table are based on the model-building data sets. All features in the final models were significant at $p < .05$ level and have variance inflation factor values less than 2, indicating negligible multicollinearity. The first human ratings on the writing fundamentals rubric were used for analyses. BA = ban ads; CG = cash for grades; SN = social networking; SL = service learning; CF = cultural fair; GG = generous gift.

First, the model *R*-squared values ranged from .36 to .48 across the six prompts, with the values being generally slightly higher for the written recommendation prompts (i.e., .45 for SL, .39 for CF, and .48 for GG) than for the argument essay prompts (i.e., .36 for BA, .39 for CG, and .36 for SN). Derived from the *R*-squared values, the correlation coefficients of the resulting predicted scores and human ratings (on writing fundamentals) ranged from .60 to .69.

Second, we noticed that certain features did not enter the final models and that the feature sets that failed to enter the models generally appeared to be consistent within a genre. For example, pauses between sentences and multiple backspacing did not enter the models for all argument essay prompts, and pauses between words did not enter the models for all written recommendation prompts. One feature (i.e., pauses between words) did not enter two of the three models in the argument essay prompts, and two features (i.e., pauses between sentences and multiple backspacing) entered the model for two of the three models in the written recommendation prompts.

Third, according to Table 7, the features total time on task and burst length in words contributed most to prediction of fundamental text production skills for all prompts in both writing genres. Two other features were also significant in predicting fundamental text production skills across all prompts: burst duration in milliseconds and single backspace events.

Finally, while we found some level of consistency within a writing genre, we also observed noticeable discrepancies between genres. For example, variation of burst duration in milliseconds accounted for considerably less variance in the argument essay prompts than in the written recommendation prompts. The opposite appeared to be true for single backspacing and pauses within words, although the contrast between the two genres was not as drastic for the burst duration feature. Finally, the total time on task appeared to have accounted for somewhat more variance in the argument essay prompts than for the written recommendation prompts.

Results for Research Question 4

RQ4: If we train a predictive model on one prompt, how well do its predictions generalize to the other prompts of the same or different genre?

Tables 8 and 9 provide information (for the argument essay and written recommendation genres respectively) about how well keystroke timing/process features generalize across prompts when used to predict human ratings on human scores for the writing fundamentals rubric. The models indicated in these tables are the same models that were developed in previous steps of the study (shown in Table 7).

Table 8 Correlation With Human Ratings (Argument Essay Genre)

Position	Form	Model based on BA	Model based on CG	Model based on SN
Within genre	BA	**.56**	.56	.57
	CG	.58	**.61**	.61
	SN	.57	.57	**.55**
Cross genre	SL	.61	.60	.61
	CF	.61	.61	.61
	GG	.64	.65	.65

Note. Values in bold represent the original model's (Table 7) performance on the base prompt on which the model was built. The first human ratings on the writing fundamentals rubric were used for analyses. BA = ban ads; CG = cash for grades; SN = social networking; SL = service learning; CF = cultural fair; GG = generous gift.

Table 9 Correlation With Human Ratings (Written Recommendation Genre)

Position	Form	Model based on SL	Model based on CF	Model based on GG
Within genre	SL	.66	.65	.65
	CF	.64	.66	.64
	GG	.68	.68	.68
Cross genre	BA	.46	.45	.46
	CG	.49	.48	.50
	SN	.45	.44	.46

Note. Values in bold represent the original model's (Table 7) performance on the base prompt on which the model was built. The first human ratings on the writing fundamentals rubric were used for analyses. BA = ban ads; CG = cash for grades; SN = social networking; SL = service learning; CF = cultural fair; GG = generous gift.

Within a genre, the models generalized rather well for all prompts in both genres. There were only small differences in the size of the correlation with human scores when the same model was applied to data from different prompts. In some cases, the models even performed better on the parallel prompts than on the base prompts themselves. For example, two of the three prompts in the argument essay genre (i.e., BA and SN) produced greater agreement for the two parallel prompts than for the base prompts. When there was a degradation in model performance from base prompts to parallel prompts, the reduction was fairly small, with the largest degradation being .05 in magnitude.

Across genres, the models trained on the argument essays generalized fully to the recommendation essay sets. When the argument models were applied to the prompts in the written recommendation genre, the models yielded correlational strength that was equal to or greater than the models' performance on the base prompts. For example, the model trained on BA responses produced correlation coefficients of .61, .61, and .64 for the three written recommendation prompts, respectively, higher than .56 for the base prompt itself. Similar results were found for the prompts CG and SN. On the other hand, the models trained on the recommendation essays did not generalize quite as well to the prompts in the argument essay genre. The models' performance was considerably less satisfactory in the argument prompts than in the base prompts in all cases. In several cases, the reduction in correlation was even greater than .20 (i.e., from SL to SN, from CF to BA and SN, and from GG to BA and SN).

If we go one step further, and examine how well the predicted scores on one prompt correlate with students' writing fundamental rubric scores in a parallel prompt, we notice discrepancies. For example, using the independent evaluation data sets, the predicted scores from the BA model associated with students' performance on CG at a correlation coefficient level of .37, but the predicted scores from the CG model associated students' performance on BA at a somewhat lower level of .30. Noticeable discrepancies were also observed in all other pairs of parallel prompts in both writing genres (Table 10). However, the size of the sample in these cases is fairly small (in many cases just above 100).

Results Synthesis

The patterns revealed in the results suggest that some of the keystroke timing and process features are particularly stable across tasks and/or particularly consistent in their relationship with writing performance in terms of fundamental text production skills.

Table 10 Correlation Coefficients of Predicted Scores With Human Ratings on a Parallel Form

	Human ratings on a different prompt taken by the same students					
	Argument essay forms			Written recommendation forms		
	BA	CG	SN	SL	CF	GG
BA	—	.37 (120)	.25 (104)			
CG	.30 (118)	—	.42 (210)			
SN	.44 (94)	.48 (223)	—			
SL				—	.47 (127)	.57 (118)
CF				.43 (121)	—	.52 (240)
GG				.42 (114)	.36 (233)	—

Note: The table correlates human ratings with predicted scores resulting from the base-prompt scoring model. Numbers in parentheses after each correlation indicate number of observations. The first human ratings on the writing fundamentals rubric were used for analyses. The cross-genre cells were empty by design because no student took two test forms from the two different genres. BA = ban ads; CG = cash for grades; SN = social networking; SL = service learning; CF = cultural fair; GG = generous gift.

- If we examine Table 5, we observe that features related to fundamental text production fluency (time on task, burst length in words, and pauses within words) have the strongest correlations across tasks, suggesting that these features reflect a relatively stable pattern of performance that is not too strongly affected by task-specific requirements. On the other hand, the features most likely to link with editing and planning behaviors (pauses between sentences, cut/paste/jump events, and backspaces) show the least consistency across tasks, suggesting more sensitivity to task requirements.
- If we examine Table 6, we observe that time on task and burst length in words have consistent positive correlations and that pauses within words have a consistent negative correlation with writing scores for fundamental text production. On the other hand, while some events potentially reflecting editing and planning (i.e., cut/paste/jump events and pauses between sentences) appear to have a consistent relationship with writing scores, other events, such as pauses between words and backspacing, do not.
- If we examine Table 7, we observe that the keystroke timing and process features accounted for a considerable amount of variance in human ratings of fundamental text production. The distribution of the reliable variance in human ratings explained by keystroke features appeared to be consistent across prompts with one drastic exception. That is, the written recommendation models rely rather more heavily on the burst duration feature. This feature is assigned much smaller weights in the argument essay models (7.2%, 10.5%, and 6.5% in the three argument essay prompts vs. 20.0%, 17.5%, and 18.1% in the three written recommendation prompts). As Tables 5 and 6 indicate, the burst duration feature is more stable in terms of cross-form correlations in the written recommendation genre. The feature has a consistent and statistically significant association with writing fundamental quality only in the written recommendation genre and has insignificant and marginal association with writing fundamental quality in the argument essay genre. The failure of the recommendation models to generalize toward argument essays (as demonstrated in Table 9 in contrast with Table 8) might therefore be primarily because of their reliance on this feature.

Note, however, that other factors may affect the difference between the two genres. The written recommendation forms were attempted primarily by seventh-grade students, and the argument essay forms were attempted primarily by eighth-grade students, and we might reasonably expect higher performance in later grades. As Table 3 shows, the mean scores ranged slightly higher for the forms administered in eighth grade. This situation suggests an additional hypothesis about the lower performance of keystroke features in predicting scores on the argument essays. It may be that more of the eighth-grade students were at higher levels of fluency, where the keystroke features would provide little information to discriminate among them. While the data we have cannot decisively support or refute this hypothesis, they suggest a potentially fruitful question for further study. Does the predictive value of process features decrease with age, as more and more students achieve the critical level of fluency for successful performance?

Discussion, Limitations, and Conclusion

In this study, we found with respect to RQ1 that some keystroke timing and process features appeared to be reasonably stable across parallel forms (Table 5). With respect to RQ2 and RQ3, we found that many of these features were correlated with and predictive of fundamental text production skill (Tables 6 and 7). With respect to RQ4, we found that models generalized well across prompts within a genre (Tables 8 and 9). In terms of the cross-genre generalization, the models for the argument essay prompts, which did not rely heavily on the burst duration feature, generalized well to the prompts in the written recommendation genre but not vice versa. Table 7 is the most striking illustration of this point, because it indicates a fairly comparable pattern of weights across prompts, while the most obvious difference between the two genres lies in the weights given to the burst duration feature.

Overall, the patterns of correlations and regression weights are consistent with the cognitive interpretations we offered earlier in this paper. The only feature with a somewhat puzzling behavior is the burst duration feature, which appears to be more stable across topics (and have a stronger correlation with writing fundamental skills) in the written recommendation genre. We are not sure how to interpret this pattern. Otherwise, it seems plausible that we could use the patterns of keystroke features as a method with which to provide additional evidence about fundamental text production skill, supplementing the information coming from human or automated scores on the final product. The models we have built indicate that middle-school students with lower overall skill in fundamental text production (as measured by a writing fundamental scoring rubric) spend less time on task and produce text less fluently, with more hesitation, whereas middle-school students with stronger fundamental text production skills write longer and more fluently while producing slightly more editing behaviors. These results are in line with theoretical expectations, as are the high correlations we observed between the writing fundamentals and content rubrics. All of these results are consistent with the causal explanation that higher fluency enables writers to devote more cognitive resources to achieving content quality goals.

It is important to note that this study has several limitations that should be taken into account when considering the conclusions offered above. First, we only tested two types of writing (argument essays and written recommendations), which are more similar to one another than the argument essay/literary analysis contrast examined by Deane (2014). We cannot, therefore, easily generalize these results to the full range of genres students are typically assigned in school. Second, the sample we collected (sixth- to ninth-grade US students) was a convenience sample that contained a very small proportion of African American or ELL students. We will need to conduct additional studies before we can draw any firm conclusions about how this limited sampling has affected our results. Third, the tests were given under low-stakes condition. Although the test administration manual provided detailed instructions for the teachers who monitored the test sessions, it was possible that students might leave in the middle of an essay task session and return later or (even more likely, judging by the time-on-task data) that students with low motivation might submit their essays early and remain in class without using their full allotted time. Such scenarios would affect the analysis of keystroke timing features, because the present design does not enable us to separate effects of motivation from limitations in text-production fluency. Finally, the keystroke features we have examined are simple summary features that indicate very little about the details of the writing process. It seems likely that a more nuanced analysis of the keystroke log data could extract richer information, and support deeper inferences, than the relatively simple summary features presented in this study.

Up to this point, the analysis we have done has been entirely at a whole-group level. It seems plausible that one could use keystroke features to explore differences between population groups, such as ELL students versus non-ELL students, or to track the growth of fundamental text production over time. In principle, the information provided by keystroke features could provide useful information for teachers at the level of individual students. For example, we could track performance longitudinally (as measured by the features predictive in our models), which could inform teachers if their students are on target for growth in fundamental text production skill for their grade level. Somewhat more speculatively, a lack of editing behavior (in combination with a low-quality text) could prompt the teacher to focus attention on editing strategies, though further studies would be needed to determine whether that is, in fact, the correct pedagogical response. More generally, given a profile of fundamental text production behavior, it becomes possible to explore how different student profiles respond to instruction, though the usefulness of such applications remains a subject for future research. To provide feedback based on process data, it would be necessary to create summary indicators and graphical representations that convert evidence from summary features into user-friendly dashboards. This is where computer–human interaction research and interface design may become highly relevant.

Notes

1 Based on the test administration design, no student took two test forms from the two different writing genres.
2 We analyzed the performance of individual raters to account for the cases where interrater agreement was unexpectedly low and identified specific individual raters who appear not to have fully complied with the directions they were given in rater training. To support future analyses, we are conducting a partial rescoring, since we believe that the low rates of agreement on the BA and SN prompts were due primarily to relatively poor performance on the part of those specific individuals.

References

Abbott, R. D., Berninger, V. W., & Fayol, M. (2010). Longitudinal relationships of levels of language in writing and between writing and reading in grades 1 to 7. *Journal of Educational Psychology*, *102*, 281–298.

Alamargot, D., & Chanquoy, L. (2001). *Studies in writing: Vol. 9. Through the models of writing*. Dordrecht, The Netherlands: Kluwer.

Alamargot, D., & Fayol, M. (2009). Modelling the development of written composition. In R. Beard, D. Myhill, J. Riley, & M. Nystrand (Eds.), *SAGE handbook of writing development* (pp. 23–47). London, England: Sage.

Almond, R., Deane, P., Quinlan, T., Wagner, M., & Sydorenko, T. (2012). *A preliminary analysis of keystroke log data from a timed writing task* (Research Report No. RR-12-23). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2012.tb02305.x

Alves, R. A., Branco, M., Castro, S. L., & Olive, T. (2012). Effects of handwriting skill, output modes, and gender on fourth graders' pauses, language bursts, fluency, and quality. In V. Berninger (Ed.), *Past, present, and future contributions of cognitive writing research to cognitive psychology* (pp. 389–402). East Sussex, England: Psychology Press.

Andersson, B., Dahl, J., Holmkvist, K., Holsanova, J., Johansson, V., Karlsson, H. … Wengelin, A. (2006). Combining keystroke logging with eye-tracking. *Writing and Digital Media*, *17*, 45–72.

Baaijen, V. M., Galbraith, D., & de Glopper, K. (2012). Keystroke analysis: Reflections on procedures and measures. *Written Communication*, *29*(3), 246–277. doi:10.1177/0741088312451108

Beauvais, C., Olive, T., & Passerault, J. M. (2011). Why are some texts good and others not? Relationship between text quality and management of the writing process. *Journal of Educational Psychology*, *103*, 415–428.

Bennett, R. E. (2011). *CBAL: Results from piloting innovative K–12 assessments* (Research Report No. RR-11-23). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2011.tb02259.x

Bennett, R. E., & Gitomer, D. H. (2009). Transforming K–12 assessment: Integrating accountability testing, formative assessment, and professional support. In C. Wyatt-Smith & J. J. Cumming (Eds.), *Educational assessment in the 21st century: Connecting theory and practice* (pp. 43–62). New York, NY: Springer.

Bereiter, C., & Scardamalia, M. (Eds.). (1987). *The psychology of written composition*. Hillsdale, NJ: Erlbaum.

Berninger, V. W. (1996). *Reading and writing acquisition: A developmental neuropsychological perspective*. Boulder, CO: Westview Press.

Berninger, V. W. (1999). Coordinating transcription and text generation in working memory during composing: Automatic and constructive processes. *Learning Disability Quarterly*, *22*, 99–112.

Breetvelt, I., van den Bergh, H., & Rijlaarsdam, G. (1994). Relations between writing processes and text quality: When and how? *Cognition and Instruction*, *12*, 103–123.

Chenoweth, N. A., & Hayes, J. R. (2003). The inner voice in writing. *Written Communication*, *20*, 99–118.

Connelly, V., Dockrell, J. E., Walter, K., & Critten, S. (2012). Predicting the quality of composition and written language bursts from oral language, spelling, and handwriting skills in children with and without specific language impairment. *Written Communication*, *29*, 278–302.

Deane, P. (2014). *Using writing process and product features to assess writing quality and explore how those features relate to other literacy tasks* (Research Report No. RR-14-03). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/ets2.12002

Deane, P., Fowles, M., Baldwin, D., & Persky, H. (2011). *The CBAL summative writing assessment: A draft eighth-grade design* (Research Memorandum No. RM-11-01). Princeton, NJ: Educational Testing Service.

Deane, P., & Quinlan, T. (2010). What automated analyses of corpora can tell us about students' writing skills. *Journal of Writing Research*, *2*, 151–177.

Deane, P., Quinlan, T., & Kostin, I. (2011). *Automated scoring within a developmental, cognitive model of writing proficiency* (Research Report No. RR-11-16). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2011.tb02252.x

Flower, L., & Hayes, J. R. (1981). A cognitive process theory of writing. *College Composition and Communication*, *32*, 365–387.

Fu, J., Chung, S., & Wise, M. (2009). *Statistical report of Fall 2009 CBAL writing tests* (Research Memorandum No. RM-13-01). Princeton, NJ: Educational Testing Service.

Graham, S., Berninger, V. W., Abbott, R. D., Abbott, S. P., & Whitaker, D. (1997). Role of mechanics in composing of elementary school students: A new methodological approach. *Journal of Educational Psychology*, *89*, 170–182.

Hayes, J. R. (2012). Modeling and remodeling writing. *Written Communication*, *29*, 369 – 388.

Hayes, J. R., & Flower, L. S. (1980). Identifying the organization of writing processes. In L. W. Gregg & E. R. Steinberg (Eds.), *Cognitive processes in writing* (pp. 3 – 30). Hillsdale, NJ: Erlbaum.

Jakobsen, A. L. (2006). Research methods in translation: Translog. In K. P. H. Sullivan & E. Lindgren (Eds.), *Studies in writing: Vol. 18. Computer keystroke logging and writing* (pp. 95 – 105). Amsterdam, The Netherlands: Elsevier.

Kaufer, D. S., Hayes, J. R., & Flower, L. (1986). Composing written sentences. *Research in the Teaching of English*, *20*, 121 – 140.

Kellogg, R. T. (2001). Competition for working memory among writing processes. *The American Journal of Psychology*, *114*, 175 – 191.

Kellogg, R. T. (2008). Training writing skills: A cognitive developmental perspective. *Journal of Writing Research*, *1*, 1 – 26.

Leijten, M., Hoste, V., Van Waes, L., Macken, L., & Van Horenbeeck, E. (2012, April). From character to word level: Enabling the linguistic analyses of Inputlog process data. In M. Piotrowski, C. Mahlow, & R. Dale (Eds.), *Proceedings of the Second Workshop on Computational Linguistics and Writing (CLW 2012). Linguistic and Cognitive Aspects of Document Creation and Document Engineering* (pp. 1 – 8). Stroudsburg, PA: Association for Computational Linguistics.

Leijten, M., & Van Waes, L. (2006). Inputlog: New perspectives on the logging of on-line writing processes in a Windows environment. In K. P. H. Sullivan & E. Lindgren (Eds.), *Studies in writing: Vol. 18. Computer key-stroke logging and writing: Methods and applications* (pp. 73 – 94). Oxford, England: Elsevier.

Leijten, M., & Van Waes, L. (2013). Keystroke logging in writing research using Inputlog to analyze and visualize writing processes. *Written Communication*, *30*, 358 – 392.

Limpo, T., & Alves, R. A. (2013). Modeling writing development: Contribution of transcription and self-regulation to Portuguese students' text generation quality. *Journal of Educational Psychology*, *105*, 401 – 413.

McCutchen, D. (1996). A capacity theory of writing: Working memory in composition. *Educational Psychology Review*, *8*, 299 – 325.

McCutchen, D. (2000). Knowledge, processing, and working memory: Implications for a theory of writing. *Educational Psychologist*, *35*, 13 – 23.

Miller, K. S. (2000). Academic writers on-line: Investigating pausing in the production of text. *Language Teaching Research*, *4*, 123 – 148.

Perl, S. (1979). The composing processes of unskilled college writers. *Research in the Teaching of English*, *13*, 317 – 336.

Strömqvist, S., Holmqvist, K., Johansson, V., Karlsson, H., & Wengelin, Å. (2006). What keystroke-logging can reveal about writing. In K. P. H. Sullivan & E. Lindgren (Eds.), *Studies in writing: Vol. 18. Computer keystroke-logging and writing* (pp. 45 – 72). Amsterdam, The Netherlands: Elsevier.

Tillema, M., van den Bergh, H., Rijlaarsdam, G., & Sanders, T. (2011). Relating self reports of writing behaviour and online task execution using a temporal model. *Metacognition and Learning*, *6*, 229 – 253.

Torrance, M., & Galbraith, D. (2006). The processing demands of writing. In C. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 67 – 80). New York, NY: Guilford Press.

van den Bergh, H., & Rijlaarsdam, G. (2001). Changes in cognitive activities during the writing process and relationships with text quality. *Educational Psychology*, *21*, 373 – 385.

van Waes, L., & Leijten, M. (2006). Logging writing processes with Inputlog. *Writing and Digital Media*, *17*, 158 – 166.

Wengelin, Å. (2006). Examining pauses in writing: Theory, methods and empirical data. In K. P. H. Sullivan & E. Lindgren (Eds.), *Studies in writing: Vol. 18. Computer keystroke-logging and writing* (pp. 107 – 130). Amsterdam, The Netherlands: Elsevier.

Zhang, M., & Deane, P. (in press). *Process features in writing*: *Internal structure and incremental value over product features*. ETS Research Report Series.

Suggested citation:

Deane, P., & Zhang, M. (2015). *Exploring the feasibility of using writing process features to assess text production skills* (ETS Research Report No. RR-15-26). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/ets2.12071

Action Editor: Beata Beigman Klebanov

Reviewers: Gary Feng and Tanner Jackson

ETS, the ETS logo, and LISTENING. LEARNING. LEADING. are registered trademarks of Educational Testing Service (ETS). CBAL is a trademark of ETS. All other trademarks are property of their respective owners.

Find other ETS-published reports by searching the ETS ReSEARCHER database at http://search.ets.org/researcher/