Listening. Learning. Leading.®

Research Report
ETS RR–15-27

Process Features in Writing: Internal Structure and Incremental Value Over Product Features

Mo Zhang

Paul Deane

December 2015

ETS Research Report Series

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

Process Features in Writing: Internal Structure and Incremental Value Over Product Features

Mo Zhang & Paul Deane

Educational Testing Service, Princeton, NJ

In educational measurement contexts, essays have been evaluated and formative feedback has been given based on the end product. In this study, we used a large sample collected from middle school students in the United States to investigate the factor structure of the writing process features gathered from keystroke logs and the association of that latent structure with the quality of the final product (i.e., the essay text). The extent to which those process factors had incremental value over product features was also examined. We extracted 29 process features using the keystroke logging engine developed at Educational Testing Service (ETS). We identified 4 factors that represent the extent of writing fluency, local word-level editing, phrasal/chunk-level editing, and planning and deliberation during writing. We found that 2 of the 4 factors — writing fluency, and planning and deliberation — significantly related to the quality of the final text, whereas the 4 factors altogether accounted for limited variance in human scores. In 1 of the 2 samples studied, the keystroke-logging fluency factor added incrementally, but only marginally, to the prediction of human ratings of text-production skills beyond product features. The limited power of the writing process features for predicting human scores and the lack of clear additional predictive value over product features are not surprising given that the human raters have no knowledge of the writing process leading to the final text and that the product features measure the basic text quality specified in the human scoring rubric. Study limitations and recommendation for future research are also provided.

Keywords *CBAL*™; writing process; keystroke logging; factor structure; incremental value

doi:10.1002/ets2.12075

Writing is a mental activity or process resulting from the interactions of multiple cognitive subprocesses that a writer implements to generate, express, and refine one's ideas while formulating a text (Berninger, 1994; Flower & Hayes, 1981; Hayes, 2012). These subprocesses, to a large extent, engage short-term working memory (WM), a temporary information store used in a wide range of cognitive tasks (McCutchen, 2000).

Research in cognition shows that text production involves at least four essential subprocesses: planning (generating and organizing ideas), translating (a linguistic operation), transcribing (motor execution), and reviewing (reading and revising the text or the plan of the text); Deane, Fowles, Baldwin, & Persky, 2011; Kellogg, 2001). These subprocesses compete with one another during composition for limited WM capacity, each subprocess plays a critical role, and none could stand alone in the completion of a writing task (Hayes, 2012; Olive, Kellogg, & Piolat, 2008).

In addition to involving multiple processes, writing is a highly recursive and interleaved activity, with continuing shifts among the different subprocesses as composition proceeds. Acknowledging the temporal nature of the writing process is critical in that, at a given stage of composition, certain activities will dominate at the expense of others (de Larios, Manchón, Murphy, & Marín, 2008; Miller, 2000). For example, planning tends to decrease while reviewing tends to increase (Kellogg, 2001, 2008). Depending on when a given subprocess occurs, it may have a different purpose and effect. In one study, the authors found that planning that occurs at the beginning of writing is more effective than planning that occurs toward the end (Breetvelt, van den Bergh, & Rijlaarsdam, 1994).

For a given task, different writers may execute strategies incurring different configurations of the subprocesses. For instance, research has indicated that experienced writers vary in their planning and revision patterns more than unskilled writers. The revision-related activities conducted by the more skilled writers mainly pertain to the meaning of the texts. This kind of revision is also referred to as conceptual revision, and it involves more extensive alteration of the content; in contrast, less skilled writers tend to show more local monitoring behavior, which is mainly characterized by corrections to punctuation, syntax, and spelling (Breetvelt et al., 1994).

Corresponding author: M. Zhang, E-mail: mzhang@ets.org

One method for studying writing process is keystroke logging (KL). KL can capture each action and its associated temporal information in a concurrent and unobtrusive way (Leijten & van Waes, 2013). The following categories of action and temporal information, either alone or in combination, can be obtained from a well-designed KL system: (a) action (type of behavior; e.g., deletion); (b) duration (time elapsed or text length that a given action covers; e.g., burst); (c) location (where in the text a given action occurs; e.g., within word); and (d) time point (when a given action occurs in the writing process; e.g., in the beginning of the composition).

KL can provide information not only about the mechanical process of the text's production, but also potentially about some of the essential underlying cognitive subprocesses cited above, including their distribution. Those subprocesses and their distribution can be evidenced by the action and temporal patterns extracted from KL.

Because these subprocesses compete for limited resources from WM, a writer must be able to effectively manage such subprocesses both globally and locally. Previous research has indicated that this management may partially explain the quality of the writing product, although the relationship may be task or writing-purpose specific (e.g., Breetvelt et al., 1994; Deane et al., 2011; Ruth & Murphy, 1988).

In addition to effectively managing subprocesses, a sufficient level of fluency in text generation is critical. In fact, managing subprocesses may be facilitated by high levels of fluency. In her review of the relationship between writing processes and working memory, McCutchen (2000) found that linguistic fluency can help reduce short-term WM load and allow the writer to more easily retrieve resources from long-term memory, leading to a higher quality final product. McCutchen's conclusions have been echoed by other researchers, who have suggested that without a certain level of fluency in lexical generation, writers cannot move beyond the constraints imposed by short-term WM (e.g., Deane, 2014; Xu & Ding, 2014). In a study of middle school students, Deane and Zhang (2015) also found that burst length, an indicator of fluency, has a strong association with essay quality.

In the current investigation, we used KL to explore the relationships among various process features, including indicators of fluency, and the association of those features with the quality of the writing product. The study employed larger participant sample sizes than typically used in KL writing research, involved a set of process features that was more extensive than previously reported in empirical studies, and investigated the extent to which those features had incremental value over automatically generated product features.

Research Questions

In this study, we investigated three research questions. The questions and their justifications are given below.

- Research Question (RQ) 1. Do the relationships among process features have a theoretically meaningful factor structure? If so, what is that structure? The interpretation and potential use of the KL process features would be more strongly justified if the empirical relations among those features can be interpreted in accordance with writing theory.
- RQ 2. Are the aggregations of the process features resulting from the factor analysis in RQ 1 related to the human ratings of text production skills in theoretically explainable ways, and if so, which factors have the strongest relations? Investigation of these relationships may help to elucidate the role of more general processes (e.g., planning) in affecting the quality of the final text.
- RQ 3. Do the most important process factors identified in RQ 2 account for variance in the human ratings of text production skills beyond that explained by features automatically derived from the final product? Human raters make their judgments based on the final product without knowledge of the processes leading to it. KL process features will be uniquely reflected in those ratings only to the extent that the writing processes affect the quality of the final text in ways that the product features do not capture. Put another way, if those KL factors do account for additional variance, it would suggest that these factors measure aspects of the human judgments of text quality not tapped by the automatically extracted product features.

Method

Participants

Data were collected from a multistate sample of sixth- to eighth-grade students who were part of a larger pilot test of a *CBAL*™ writing assessment in 2013. Schools were recruited for participation on a voluntary basis. Of 1,105 students

who took the assessment form used in this study, 34.6% were female, 35.6% were male, and 29.8% did not provide that information. In terms of grade level, 7.9% were sixth graders, 50.8% were seventh graders, 11.6% were eighth graders, and the rest (29.8%) did not provide grade information. The majority of students (∼58%) were English native speakers (or fluent in English upon school entry), roughly 4% were former or current English-language learners, and nearly 38% did not give relevant information. Finally, 47.9% indicated that their race/ethnicity was Caucasian; 15.9% Hispanic; 6.4% Native American, Asian, African American, Hawaiian/Pacific Islander, or mixed race; and 29.8% did not report this information.

Of the 1,105 participants, 831 provided data suitable for this study's analyses. The demographic distribution of the final data set was comparable to that described above. The reasons for exclusion are described in the data analyses section below.

Instrument

Test takers took the CBAL Generous Gift assessment form. They were asked to evaluate two proposals on how to spend a large sum of money provided by a generous donor to a school and to use explicit criteria in writing an essay recommending one proposal over another. The writing purpose of the culminating essay task is best described as policy recommendation.

By design, the essay task was given at the end of the assessment. This design allows test takers to gain a sufficient level of familiarity with the topic from a series of lead-in tasks. Each test taker was given 45 minutes to complete the essay task, which is the focus of this study.

All essays were scored by at least two human raters on a rubric evaluating basic text quality. A rater provides a score on a scale ranging from integer 0 to 5 or assigns a condition code of 9 for a blank response (see the Appendix for the scoring rubric). During scoring, if the two raters disagreed by more than two points, a third rater was asked to adjudicate. The interrater agreement between the first and second raters was .63 for quadratically weighted kappa, .67 for Pearson correlation coefficient, 51% for exact percentage agreement, and 97% for one-point adjacent percentage agreement. This interrater agreement was deemed adequate for the purpose of the current analyses. In CBAL writing assessment, raters use a second rubric to evaluate higher-level writing skills such as quality of ideas and argumentation. In this study, we were only concerned with the scoring rubric related to writing fundamentals and not with this second rubric.

Process and Product Features

Keystroke logs on the essay task for all 1,105 students were collected. Twenty-nine KL process features were extracted using the KL engine developed at Educational Testing Service (ETS).

Some of the extracted features are intended as indicators of fluency. These features are primarily based on pause patterns (or conversely, a burst of text production). Examples include the median value of the longest within-word pause time across all words, indicating the extent to which there is a clear hesitation (or lack of burst); the median value of the interkey interval pause time between the first and second characters in a word across all words, implicating an overall word-level keyboarding fluency; and the median value of between-word pause time across all pairs of words, suggesting general typing fluency.

Some other KL process features measure the extent of editing and revision. These features include the proportion of corrected or uncorrected typos over the total number of words and the proportion of the total time on task spent on multiword deletion. Still other KL process features intend to provide measures of the extent of planning and deliberation. Examples include the proportion of time spent at the start of phrasal bursts and the median value of pause length across all sentences junctures. Each of these measures might indicate the time devoted to planning and deliberation in between bursts and sentences, respectively. Table 1 shows the KL process features analyzed in this study.

The product features on all submitted essays were obtained using the *e-rater*® automated scoring system developed at ETS (ETS, 2014). Ten product features that measure vocabulary complexity, essay organization, accuracy of grammar and mechanics, and style in terms of sentence variety and word use were extracted for each response. The e-rater product features are listed in Table 2.

Data Analyses

Before conducting analyses, several steps were taken to ensure the quality of the process and product data. First, there were 114 KLs where it appeared that the test takers had closed and reopened the browser while taking the test, resulting

Table 1 Keystroke Logging (KL) Process Features

Process feature	Description
CharsInMultiWordDeletion	Extent to which deletion of characters occurs in the process of deleting multiple words
CorrectedTypo	Extent to which correction of mistyped words occurs
DeletedChararacter	Extent to which deletion of characters occurs
EditedChunk	Extent to which deleted text is replaced with edited text of similar content
EndSentencePunctuationPause	Extent to which pauses occur at the juncture of sentences, which may indicate planning and deliberation
EventsAfterLastCharacter	Extent to which editing of any kind occurs
InSentencePunctuationPause	Extent to which pauses occur at a sentence-internal punctuation mark, which may reflect sentence-level planning
InterkeyPause	Extent to which pauses occur between keystrokes, suggesting general typing fluency
LongJump	Extent to which jumps to different areas in the text occur
MajorEdit	Extent to which words are edited beyond minor correction
MinorEdit	Extent to which words are edited to make only minor corrections
MultiWordDeletion	Extent to which multiword text deletion occurs
MultiWordEditTime	Extent of time spent in deleting multiple words
NewContentChunk	Extent to which deleted text is replaced with edited text with new content
PhrasalBurst	Extent to which long sequences of words are produced without interruption, possibly reflecting planning or deliberation
PreJumpPause	Extent to which pauses occur before jumping to a different part of the document, possibly suggesting planning and deliberation
RetypedChunk	Extent to which deleted text is replaced with essentially the same text
StartTime	Extent to which a pause occurs prior to beginning writing, possibly reflecting planning and deliberation
TimeSpentAtPhrasalBurst	Extent to which pauses occur at the beginning of a string of fluent text production, possibly suggesting planning and deliberation
TimeSpentBetweenPhrasalBurst	Extent to which pauses occur between strings of fluent text production, possibly suggesting planning and deliberation
TypoCorrectedChunk	Extent to which text is replaced with edited text that differs primarily in minor spelling correction
UncorrectedSpelling	Extent to which a spelling error occurs that is not corrected before another unrelated action is taken
WordChoice	Extent to which words are edited to produce completely different words, possibly suggesting deliberation about word choice
WordChoiceEventPause	Extent to which pauses occur when replacing words with different words, possibly reflecting deliberation over word choice
WordEditingPause	Extent to which pauses occur within words during text editing
WordFinalPause	Extent to which pauses occur just before typing the last character in a word, a measure of general typing fluency
WordInitialPause	Extent to which pauses occur just after typing the first character in a word, which could reflect on the one hand, general typing fluency, or on the other hand, deliberation for word choice, retrieving spelling, or planning keystroke sequences for a word
WordInternalPause	Extent to which pauses occur within words during text production
WordSpacePause	Extent to which pauses occur in between words, suggesting general typing fluency

in a partial loss of data. Those responses were excluded from the final data set. Second, we excluded 70 responses that received human ratings of 0 or 9, because such a rating indicates some kind of aberrancy such as a non-English, random keystroke, or empty response. Third, a very short KL, or an essay composed within an extremely short time, contains limited information, and the process features extracted from the log are unlikely to be reliable measurements of the writing process. Hence, we eliminated a small number of essays with fewer than 25 words and/or two sentences and essays where the total time on task was less than 3 minutes. Fourth, in some cases, the ratio of number of words to time on task suggested that the entire response was pasted in with very little time spent on its composition; those responses were excluded as well. Finally, we excluded responses that were flagged by e-rater as problematic in terms of being off-topic, having unrecognizable organization, or having an excessive number of mechanical errors. All results reported are based on the final cleaned data set containing essay responses by 831 test takers.

Table 2 e-rater Product Features

Product feature	Description
Grammar	Absence of errors in pronouns, run-ons, missing possessives, etc.
Mechanics	Absence of errors in capitalization, punctuation, commas, hyphens, etc.
Style	Absence of errors in repetition of words, inappropriate words, etc.
Usage	Absence of errors in missing/wrong articles, nonstandard verbs, etc.
Collocation/preposition	Extent of the correct choice and usage of juxtaposition of words
Organization	Average length of discourse units
Development	Number of discourse units
Word frequency	Median value of word frequency measured by standard frequency index
Word length	Average word length
Sentence variety	Diversity of sentence structure

We randomly divided the sample into two sets: one with 500 essays (Sample A) and the other with 331 essays (Sample B). Sample A was used to investigate all research questions, and Sample B was saved for cross-validation.

To address RQ 1, we first ran a principal factor analysis on all 29 writing process features using Sample A, retaining factors with eigenvalues greater than 1. We rotated the factors using promax oblique rotation and examined the rotated factor pattern and scree plot. This analysis, along with substantive analysis of the factor pattern, suggested a further reduction in the number of dimensions. Loadings of greater than .30 were identified and used in the interpretation of the resulting factors.

To address RQ 2, we first computed KL factor scores for all responses from Sample A. Using multiple linear regression, we regressed the adjudicated human scores on the factor scores and examined the significance of each factor in explaining the reliable variance in human ratings.[1] For subsequent analysis in RQ 3, we listed the factors in descending order based on their significance levels and relative weights in the regression model.

To address RQ 3, we ran a hierarchical multiple regression by first entering the e-rater product features as a block and then each of the four factors sequentially in steps based on the ordering resulting from RQ 2. We then examined whether the adjusted R-squared value was significantly increased in each step.

We conducted cross-validation on Sample B. We first computed the factor scores for all the essays in Sample B using the coefficients resulting from the factor analyses in Sample A. We then replicated the multiple linear regression and hierarchical regression in Sample B to see whether the same set of factors significantly contributed to the reliable variance in human ratings and to determine whether the incremental value was the same as identified in Sample A.

Results

Table 3 shows the summary statistics for the KL process features, the e-rater product features, and the human ratings for Sample A and Sample B. As can be seen, the samples appear quite comparable. The test takers' text-production skills, as evidenced by the human-rating distributions, are almost identical across samples (Sample A: Mean = 2.51, $SD = 0.83$; Sample B: Mean = 2.50, $SD = 0.79$). The writing process feature distributions and text characteristics of the final product are also similar between samples. It is noteworthy that a few KL process features, particularly the ones related to word- and chunk-level editing behavior, have means and standard deviations of nearly zero. This result indicates that, in general, the number of occurrences was very low, or the duration of the actions was short. The most extreme examples include the proportion of major edits, proportion of retyped chunks, proportion of typo-corrected chunks, proportion of new content chunks, and proportion of multiword deletions.

Results for RQ 1

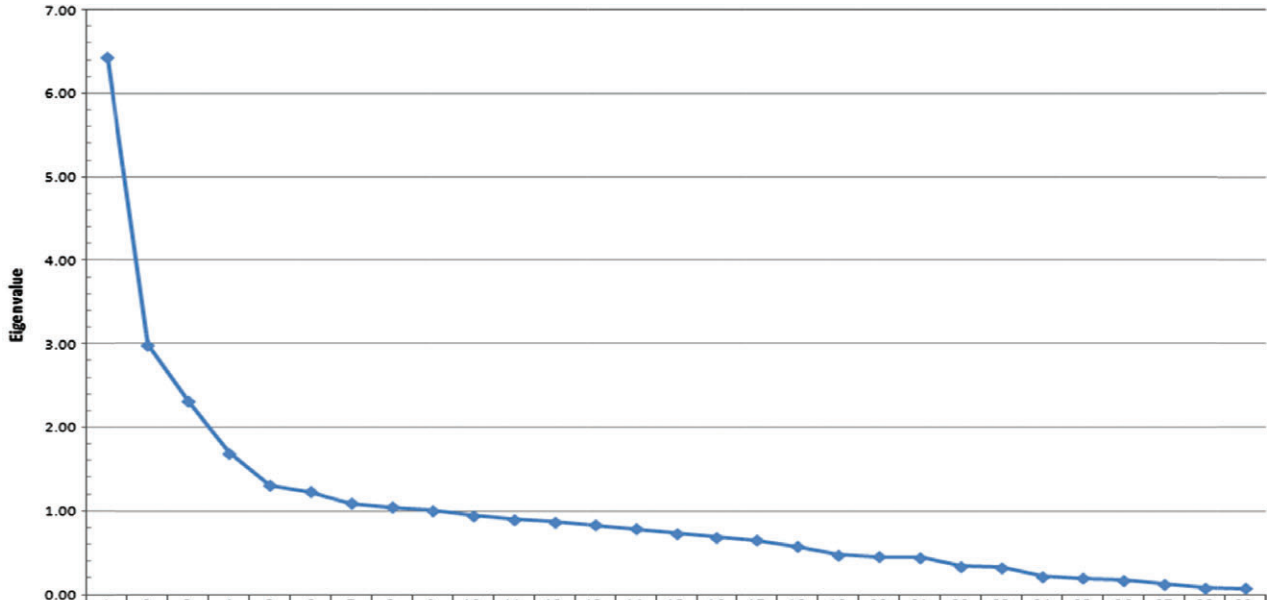
In the first research question, we ask whether the relationships among KL process features have a theoretically meaningful factor structure and, if so, what that structure would be. Nine factors were retained based on the eigenvalue criterion of 1.0. Substantive analyses of the rotated factor pattern, along with visual analyses of the scree plot (Figure 1), suggested that a four-factor structure would provide a more theoretically meaningful and parsimonious solution.[2] A second principal factor analysis was therefore conducted, specifying the number of factors to be four.

Table 3 Distributions of the Human Scores, Keystroke Logging (KL) Process Features, and Product Features

	Sample A ($n = 500$)		Sample B ($n = 331$)	
Variable	Mean	*SD*	Mean	*SD*
Adjudicated human scores	2.514	0.833	2.503	0.789
CharsInMultiWordDeletion	0.069	0.069	0.064	0.062
CorrectedTypo	0.007	0.004	0.007	0.004
DeletedChararacter	0.137	0.077	0.135	0.079
EditedChunk	0.001	0.001	0.001	0.001
EndSentencePunctuationPause	14.530	21.981	13.869	19.291
EventsAfterLastCharacter	0.956	0.072	0.961	0.068
InSentencePunctuationPause	8.172	13.257	8.823	18.018
InterkeyPause	0.295	0.092	0.300	0.095
LongJump	0.013	0.034	0.014	0.038
MajorEdit	0.000	0.001	0.000	0.001
MinorEdit	0.017	0.008	0.017	0.008
MultiWordDeletion	0.006	0.004	0.005	0.004
MultiWordEditTime	0.050	0.041	0.046	0.039
NewContentChunk	0.002	0.001	0.001	0.001
PhrasalBurst	16.052	6.769	15.381	7.005
PreJumpPause	6.804	6.903	6.673	8.471
RetypedChunk	0.000	0.000	0.000	0.000
StartTime	0.140	0.091	0.149	0.096
TimeSpentAtPhrasalBurst	0.281	0.103	0.279	0.099
TimeSpentBetweenPhrasalBurst	0.137	0.106	0.135	0.126
TypoCorrectedChunk	0.000	0.000	0.000	0.000
UncorrectedSpelling	0.013	0.006	0.014	0.007
WordChoice	0.005	0.003	0.005	0.003
WordChoiceEventPause	2.560	4.529	2.884	7.993
WordEditingPause	0.993	0.601	0.972	0.490
WordFinalPause	0.234	0.085	0.238	0.091
WordInitialPause	0.577	0.286	0.619	0.348
WordInternalPause	0.409	0.165	0.419	0.188
WordSpacePause	0.220	0.065	0.219	0.072
Grammar	−0.087	0.038	−0.090	0.035
Mechanics	−0.218	0.092	−0.225	0.090
Style	−0.350	0.132	−0.362	0.132
Usage	−0.086	0.061	−0.082	0.059
Collocation/Preposition	0.559	0.143	0.548	0.146
Organization	1.479	0.502	1.466	0.517
Development	3.650	0.373	3.626	0.383
Word frequency	−63.471	2.678	−63.604	2.877
Word length	4.381	0.307	4.369	0.323
Sentence variety	2.779	0.577	2.704	0.547

Note: Variables are grouped by horizontal rules: adjudicated human scores, keystroke logging (KL) process features, and product features, respectively.

This four-factor structure can be interpreted in the following way. Factor 1 is dominated by features measuring general fluency in text production (Table 4). For example, the two features that loaded most highly on this factor were the median pause time between keystrokes (.927) and the median maximum pause time within words (.919). Factor 2 is dominated by features measuring the extent of phrasal and chunk-level editing. The two features that loaded most highly on this factor were the proportion of characters that were deleted as part of multiword deletions (.876) and the proportion of multiword deletions (.846). Factor 3 is dominated by features reflecting local editing processes, involving mainly spelling error correction. Here, the proportion of minor edits (.888) and the proportion of corrected typos (.814) had the highest loadings. Finally, Factor 4 is dominated by features suggesting the extent of planning and deliberation during the writing process. The proportion of time spent before phrasal bursts (.733) and between phrasal bursts (.664) loaded the highest on this factor.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
Eigenvalue	6.44	2.99	2.31	1.70	1.30	1.22	1.09	1.04	1.01	0.95	0.90	0.87	0.83	0.79	0.74	0.69	0.65	0.57	0.48	0.45	0.44	0.34	0.32	0.22	0.20	0.18	0.13	0.08	0.07
Proportion	22.2%	10.3%	8.0%	5.9%	4.5%	4.2%	3.8%	3.6%	3.5%	3.3%	3.1%	3.0%	2.9%	2.7%	2.5%	2.4%	2.3%	2.0%	1.6%	1.6%	1.5%	1.2%	1.1%	0.7%	0.7%	0.6%	0.5%	0.3%	0.3%

Figure 1 Scree plot from principal factor analysis of 29 keystroke logging (KL) process features.

The correlations among the four factors are presented in Table 5. As the table shows, the factors are only weakly related to one another, suggesting that the factors provide largely unique information.

Results for RQ 2

In RQ 2, we address to what extent the aggregation of the process features relates to the quality of the final text as evaluated by human raters. The regression was statistically significant ($p < .0001$) with a model R-squared value of .19.

Table 6 shows the results from regressing human ratings of text production skills on the factor scores coming from the analyses conducted in RQ 1. Factor 1 (general fluency) was the most statistically significant factor ($p < .0001$) and explains the overwhelming majority of the reliable variance in human ratings of text quality (68%). The direction of the parameter estimate is also consistent with the theoretical proposition underlying how the features are computed. That is, the more disrupted the writing is, the lower the human score tends to be on the text-production-skills rubric. Factor 4 (planning and deliberation) was also statistically significant, accounting for nearly 20% of the reliable variance in human ratings. Here again, the direction of the parameter estimate fits with expectation in that less planning and deliberation led to generally lower human scores. In contrast, Factor 2 (phrasal/chunk-level editing) and Factor 3 (local word-level editing) exhibit limited and nonsignificant contributions to explaining the reliable variance in the human ratings.

For cross-validation, we used the coefficients resulting from the four-factor solution in Sample A to compute the factor scores for Sample B. Human ratings of text quality from Sample B were then regressed on these factor scores. The model was again statistically significant ($p < .0001$) with a similar R-squared value of .21.

Regression results showed the same pattern as for Sample A; that is, Factor 1 and Factor 4 were both statistically significant predictors (Table 7). Moreover, the distribution of variance accounted for by those two factors in cross-validation was almost exactly the same as in Sample A. Similarly, the two editing-related factors (Factor 2 and Factor 3) were not statistically significant, accounting for limited reliable variance in human ratings.

Results for RQ 3

In RQ 3, we investigate whether the most important KL process factors based on RQ 2 accounted for variance in the human ratings of text production skills beyond that explained by product features. We ran a hierarchical regression by first regressing the human ratings on the product features only and then entered Factor 1, Factor 4, Factor 2, and Factor 3

Table 4 Rotated Four-Factor Pattern From Principal Factor Analysis (Sample A; $n = 500$)

KL process features	Factor 1	Factor 2	Factor 3	Factor 4
InterkeyPause	**.927**	−.130	−.028	−.069
WordInternalPause	**.919**	−.158	−.044	.097
WordFinalPause	**.894**	−.161	−.157	−.017
WordInitialPause	**.814**	−.165	−.064	.239
WordSpacePause	**.789**	−.079	−.127	−.061
WordEditingPause	**.583**	−.175	**−.409**	.138
CharsInMultiWordDeletion	−.097	**.876**	−.139	.081
MultiWordDeletion	−.158	**.846**	.292	−.004
MultiWordEditTime	−.137	**.843**	−.010	.013
DeletedChararacter	−.115	**.772**	.135	.218
NewContentChunk	−.091	**.748**	.106	−.057
EditedChunk	−.150	**.540**	**.349**	−.039
MinorEdit	−.068	.037	**.888**	−.064
CorrectedTypo	−.091	.139	**.814**	−.081
WordChoice	−.215	.261	**.515**	−.143
UncorrectedSpelling	.183	.061	**.387**	.069
TimeSpentAtPhrasalBurst	.124	−.177	−.035	**.733**
TimeSpentBetweenPhrasalBurst	−.040	−.086	−.115	**.664**
EndSentencePunctuationPause	−.027	−.039	−.088	**.467**
StartTime	.023	−141	−020	**−.464**
PhrasalBurst	**−.352**	.110	−.027	**−.415**
EventsAfterLastCharacter	−.036	−.210	.066	**−.369**
InSentencePunctuationPause	.009	.044	−.007	**.360**
LongJump	.082	−.074	**−.302**	.163
WordChoiceEventPause	.229	−.093	−.174	.244
TypoCorrectedChunk	−.196	.264	.117	−.070
MajorEdit	−.170	.017	.241	−.038
PreJumpPause	.098	.132	.088	.190
RetypedChunk	−.063	−.016	.093	.044

Note: KL = keystroke logging. Boldface values have an absolute loading of .30 or above. The proportions of variance accounted for uniquely were 34.0%, 29.8%, 18.9%, and 17.3% for Factors 1 – 4, respectively.

Table 5 Interfactor Correlations (Sample A; $n = 500$)

	Factor interpretation	Factor 1	Factor 2	Factor 3
Factor 1	General fluency	—		
Factor 2	Phrasal/chunk-level editing	−.31	—	
Factor 3	Local word-level editing	−.22	.24	—
Factor 4	Planning and deliberation	.17	−.07	−.19

Note: For general fluency and local word-level editing, higher scores indicate lower levels.

in sequential steps based on the ordering of predicting the reliable variance in human ratings (RQ 2). We next conducted significance test on the incremental value of the adjusted *R*-squared at each step. The process was implemented in Sample A and then replicated in Sample B.

Table 8 provides the *R*-squared and adjusted *R*-squared values resulting from each step. As the table shows, there is a statistically significant, but small, improvement in adjusted *R*-squared (from .709 to .714) by adding Factor 1 (writing fluency) to the product features, $F(1, 500) = 7.84$, $p < .001$. None of the other factors had statistically significant incremental value. On cross-validation, none of the factors added incrementally to the product features at statistically significant levels.

Discussion

In this study, we investigated the factor structure of KL writing process features and the association of that latent structure with the quality of the final text. The extent to which those process factors had incremental value over product features

Table 6 Regression of Human Ratings of Text Quality on Keystroke Logging (KL) Factors (Sample A; $n = 500$)

Predictor	Factor interpretation	Parameter estimate	*t* value	*p* value	Relative weights
Intercept		2.51	74.78	<.0001	0.0%
Factor 1	General fluency	−0.35	−9.76	<.0001	68.0%
Factor 2	Phrasal/chunk-level editing	0.05	1.32	.19	9.2%
Factor 3	Local word-level editing	−0.02	−0.49	.63	3.3%
Factor 4	Planning and deliberation	0.10	2.93	<.001	19.5%

Note: For general fluency and local word-level editing, higher scores indicate lower levels. Relative weights add to 100%. First order correlations with human ratings for Factors 1–4 were − .42, .17, .06, and .05, respectively, with only the values for Factors 1 and 2 statistically significant at the $p < .0001$ level.

Table 7 Cross-Validation of Regressing Human Ratings of Text Quality on Keystroke Logging (KL) Factors (Sample B; $n = 331$)

Predictor	Factor interpretation	Parameter estimate	*t* value	*p* value	Relative weights
Intercept		2.50	64.63	<.0001	0.0%
Factor 1	General fluency	−0.36	−8.69	<.0001	65.9%
Factor 2	Phrasal/chunk-level editing	0.03	0.64	.52	5.4%
Factor 3	Local word-level editing	−0.03	−0.71	.48	4.9%
Factor 4	Planning and deliberation	0.13	3.26	.001	23.4%

Note: For general fluency and local word-level editing, higher scores indicate lower levels. Relative weights add to 100%. First order correlations with human ratings for Factors 1–4 were − .43, .14, .06, and .10, respectively, with the value for Factor 1 statistically significant at the $p < .0001$ level and Factor 2 at $p < .05$ level.

Table 8 Hierarchical Regression of Human Ratings on Product and Keystroke Logging (KL) Process Features

	Sample A ($n = 500$)		Cross-validation: Sample B ($n = 331$)	
Feature set	*R*-squared	Adjusted *R*-squared	*R*-squared	Adjusted *R*-squared
Product	.715	.709	.694	.684
Product + F1	.720	**.714**	.695	.685
Product + F1 + F4	.720	.714	.695	.684
Product + F1 + F4 + F2	.721	.713	.696	.683
Product + F1 + F4 + F2 + F3	.721	.713	.696	.682

Note: F1 = Factor 1; F2 = Factor 2; F3 = Factor 3; F4 = Factor 4. Product feature set is given in Table 2. Boldface indicates statistically significant incremental value at $p < .01$ from the previous step.

was also examined. Among other things, the sizes of the examinee sample of more than 800 middle school students, and of the set of 29 KL process features, exceed those previously reported in KL writing research (e.g., Chukharev-Hudilainen, 2014; de Larios et al., 2008; Miller, 2000).

Factor analysis results produced a theoretically interpretable structure, with the overwhelming majority of the process features (21 of 29) loading noticeably on only a single factor. Four factors were identified that represent the extent of writing fluency, local word-level editing, phrasal/chunk-level editing, and planning and deliberation during writing.

The four factors in combination accounted for limited variance in human scores. This finding is not surprising, given that human scores are based on evaluation of the writing product. Human raters have no knowledge of the process leading to the essay other than what they can glean from the quality of the final text. Nonetheless, two of the four factors — writing fluency, and planning and deliberation — significantly related to the quality of the final text. This result was replicated across two independent samples.

The last study finding was that the KL fluency factor added incrementally, but only marginally in one sample — and not at all in the second sample — to the prediction of human ratings of text production skills beyond automatically extracted product features. Of note is that the product features concentrate on writing basics (e.g., grammar, mechanics, vocabulary sophistication) that are well aligned with the human scoring rubric. As such, it is not surprising that the KL process features did not have incremental value over the product features in predicting the human ratings on this basic text-quality rubric.

Because of their substantive focus, these product features might not be expected to be as successful in predicting human ratings of higher-level text characteristics (e.g., the quality of ideas, strength of argument, and factual accuracy). For ratings of higher-level characteristics, it is possible that KL process features, especially the ones related to planning and deliberation, and to phrasal/chunk-level editing, might add predictive value beyond the product features.

This study had several limitations. First, the sample was a voluntary one, and as such results are best generalized to students with similar characteristics. Second, students took the test under low-stakes conditions, which might have affected motivation to revise. For example, there were very few instances of certain types of editing behavior as measured by the KL features. This low level of some types of editing behavior, in turn, might have affected the results (e.g., limiting the predictive value of editing-related factors). Finally, the data were collected from a single essay-writing session, which encourages a first-draft response offering limited opportunity for editing and revising.

As for future studies of the use of KL in writing assessment, our first recommendation is to replicate the findings with more tasks administered to similar and different examinee populations. That replication might employ alternative data reduction techniques, such confirmatory factor analysis, or other methods for aggregating features in theoretically meaningful and practically useful ways. A second recommendation is to investigate the functioning of the KL features (as individual features or in combination) across different demographic groups, such as English-language learners versus non-English-language learners. A third recommendation is to study the association between KL features and text quality at different stages of the writing process. Additionally, it is worth exploring whether and how the information collected from the writing process, in conjunction with automatically extracted product characteristics, can be used to assist teaching and learning in the classroom. For example, a lack of editing behavior, in combination with a disjointed text, could prompt the teacher to focus attention on editing strategies. Such a richer writing profile generated from both process and product information also has potential to better track and represent student growth.

Acknowledgments

We thank Shelby Haberman, André Rupp, and Randy Bennett for providing technical guidance on the study design and analysis. We also thank the editors and reviewers for their suggestions on previous versions of this manuscript. Many ETS colleagues have contributed to the participant recruitment and to scoring of the essay tasks. We would like to express particular gratitude to Lynn Zaback, Lauren Phelps, Eric Maywar, and Mary Fowles in providing leadership in those efforts.

Notes

1 The adjudicated human scores were the average of the first two human ratings in the absence of a third rater and the average of the two closest ratings among the three ratings if a third score was available. In cases where the three scores were discrepant by more than two points from one another, we used only the middle score.

2 We also examined a five-factor solution. Aside from being more parsimonious, the four-factor solution was determined to be more interpretable.

References

Berninger, V. W. (1994). *Reading and writing acquisition: A developmental neuropsychological perspective*. Madison, WI: Brown & Benchmark.

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.

Chukharev-Hudilainen, E. (2014). Pauses in spontaneous written communication: A keystroke logging study. *Journal of Writing Research*, *6*, 61–84.

de Larios, J. R., Manchón, R., Murphy, L., & Marín, J. (2008). The foreign language writer's strategic behavior in the allocation of time to writing processes. *Journal of Second Language Writing*, *17*, 30–47.

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

Educational Testing Service. (2014). *About the e-rater scoring engine*. Retrieved from https://www.ets.org/erater/about

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

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

Kellogg, R. T. (2001). Competition for working memory among writing processes. *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., & van Waes, L. (2013). Keystroke logging in writing research using Inputlog to analyze and visualize writing processes. *Written Communication*, *30*, 358 – 392.

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.

Olive, T., Kellogg, R. T., & Piolat, A. (2008). Verbal, visual, and special working memory demands during text composition. *Applied Psycholinguistics*, *29*, 669 – 687.

Ruth, L., & Murphy, S. (1988). *Designing writing tasks for the assessment of writing*. Norwood, NJ: Ablex.

Xu, C., & Ding, Y. (2014). An exploratory study of pauses in computer-assisted EFL writing. *Language Learning & Technology*, *18*, 80 – 96.

Appendix

CBAL Generic Scoring Guide: Discourse-Level Features in a Multiparagraph Text

EXEMPLARY (5)

An EXEMPLARY response meets all of the requirements for a score of 4 but *distinguishes itself by skillful use of language*, *precise expression of ideas*, *effective sentence structure*, *and/or effective organization*, which work together to control the flow of ideas and enhance the reader's ease of comprehension.

CLEARLY COMPETENT (4)

A CLEARLY COMPETENT response typically displays the following characteristics:

- It is adequately structured.
  - Overall, the response is clearly and appropriately organized for the task.
  - Clusters of related ideas are grouped appropriately and divided into sections and paragraphs as needed.
  - Transitions between groups of ideas are signaled appropriately.
- It is coherent.
  - Most new ideas are introduced appropriately.
  - The sequence of sentences leads the reader from one idea to the next with few disorienting gaps or shifts in focus.
  - Connections within and across sentences are made clear where needed by the use of pronouns, conjunctions, subordination, and so on.
- It is adequately phrased.
  - Ideas are expressed clearly and concisely.
  - Word choice demonstrates command of an adequate range of vocabulary.
  - Sentences are varied appropriately in length and structure to control focus and emphasis.
- It displays adequate control of Standard Written English.
  - Grammar and usage follow SWE conventions, but there may be minor errors.
  - Spelling, punctuation, and capitalization follow SWE conventions, but there may be minor errors.

DEVELOPING HIGH (3)

A response in this category displays some competence but differs from Clearly Competent responses in at least one important way, including *limited development*; *inconsistencies in organization*; *failure to break paragraphs appropriately*; *occasional tangents*; *abrupt transitions*; *wordiness*; *occasionally unclear phrasing*; *little sentence variety*; *frequent and distracting errors in Standard Written English*; *or relies noticeably on language from the source material*.

DEVELOPING LOW (2)

A response in this category differs from Developing High responses because it displays serious problems such as *marked underdevelopment*; *disjointed*, *list-like organization*; *paragraphs that proceed in an additive way without a clear overall focus*; *frequent lapses in cross-sentence coherence*; *unclear phrasing*; *excessively simple and repetitive sentence patterns*; *inaccurate word choices*; *errors in Standard Written English that often interfere with meaning*; *or relies substantially on language from the source material*.

MINIMAL (1)

A response in this category differs from Developing Low responses because of serious failures such as *extreme brevity*; *a fundamental lack of organization*; *confusing and often incoherent phrasing*; *little control of Standard Written English*; *or can barely develop or express ideas without relying on the source material*.

NO CREDIT (0): *Not enough of the student's own writing for surface-level features to be judged*; *Blank*; *not written in English*; *completely off-topic*; *or random keystrokes*.

OMIT (9): *Blank.*

Suggested citation:

Zhang, M., & Deane, P. (2015). *Process features in writing*: *Internal structure and incremental value over product features* (Research Report No. RR-15-27). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/ets2.12075

Action Editor: Beata Beigman Klebanov

Reviewers: Frank Williams and JiangangHao