Research Report

Analysis of Keystroke Sequences in Writing Logs

ETS RR–19-11

Mengxiao Zhu
Mo Zhang
Paul Deane

December 2019

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

Analysis of Keystroke Sequences in Writing Logs

Mengxiao Zhu, Mo Zhang, & Paul Deane

Educational Testing Service, Princeton, NJ

The research on using event logs and item response time to study test-taking processes is rapidly growing in the field of educational measurement. In this study, we analyzed the keystroke logs collected from 761 middle school students in the United States as they completed a persuasive writing task. Seven variables were extracted from the keystroke logs and compared with different score and gender groups. Group comparisons were also made using methodologies borrowed from sequence mining. Students' composition strategies over the course of the writing process were also investigated. The findings of this study have implications for gaining deeper understanding of observed group differences and for designing interventions to close the achievement gaps among population groups.

Keywords Keystroke log; sequence analysis; writing assessment

doi:10.1002/ets2.12247

The research on using event logs and item response time to study test-taking processes is rapidly growing in the field of educational measurement. For instance, researchers have used timing and process data to study test-taking strategies (Lee & Haberman, 2016), guessing behavior (Guo et al., 2016), test speededness (Ranger, Kuhn, & Gaviria, 2015), and item parameter estimation (Van der Maas, Molenaar, Maris, Kievit, & Borsboom, 2011). The *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 2014) treats the response process as a kind of validity evidence in the process of assessment development and validation. Two recently published books have been dedicated to this topic: (Ercikan & Pellegrino, 2017 and Zumbo & Hubley, 2017). These timing and process data, such as the click stream during an assessment, call for new methodologies and conceptualizations to analyze process-based evidence that is not traditionally used or readily available in psychometric analyses and present new challenges to investigators. Part of the complexity in analyzing such data stems from making meaningful interpretations and valid inferences about the cognitive processes undertaken between a task and a response, which further involves disaggregating the cognitive processes from other factors such as emotions and motivation (Leighton, Tang, & Guo, 2017). Data management, as another critical component for process data-based research and analyses, can also be a challenge faced by researchers (Hao, Smith, Mislevy, von Davier, & Bauer, 2016). Compared to a single score or a final response for an item, process data contain far more information that should be structured in a way that can meet the analysis and validation needs.

In this study, we analyzed a unique type of process data that has not previously been discussed extensively in educational assessment — the keystroke logs collected during a writing task. For a keystroke log, the content, mechanical, and temporal information for all the key presses during the text-generation process are recorded. One advantage of keystroke logging is that, compared to other data collection methods (e.g., think-aloud and eye-tracking), the data collection mechanism is relatively less intrusive. The data are collected in the background without inserting any obvious interference with the writer's performance or thinking process (Leijten & Van Waes, 2013).

Despite a scarcity of the use of keystroke logs in the assessment context, there is an extensive and growing literature on the use of keystroke logs in writing research for a variety of goals (Van Waes, Leijten, & Van Weijen, 2009). Researchers have used keystroke logs to analyze composition strategies (Xu & Ding, 2014), genre effects (Beauvais, Olive, & Passerault, 2011), and transcription skills (Grabowski, 2008), to name a few. Researchers have also used keystroke logs to compare writing skills between native and nonnative speakers (Miller, 2000; Roca de Larios, Manchón, Murphy, & Marín, 2008). Writing research in the contexts of spontaneous communication (Chukharev-Hudilainen, 2014), professional writing (Leijten, Van Waes, Schriver, & Hayes, 2013), and language translation (Dragsted & Carl, 2013) has also benefited from the availability of keystroke logs. Research on writing processes in the field of educational assessment is relatively sparse,

Corresponding author: M. Zhu, E-mail: mzhu@ets.org

but there are a few exceptions (e.g., Allen et al., 2016; Deane & Zhang, 2015; Guo, Deane, van Rijn, Zhang, & Bennett, 2018; Zumbo & Hubley, 2017) that provide empirical and theoretical support for large-scale standardized writing assessments to collect, analyze, and report writing process results.

In this study, we analyzed the keystroke logs collected in a summative writing assessment and compared the text production processes between different proficiency groups and the two gender groups, males and females. The work was conducted under a larger research goal of closing the achievement gaps among population groups. It is important to discover ways to improve the skill development that fosters successes, especially for the underperforming student groups. We will not be able to fully address this large topic in this single study, but we do aim to illustrate how analyses of writing processes with the use of keystroke logs can help identify strong and weak areas and address the specific needs of underperforming groups. Regarding the comparison of the two gender groups, there is ample evidence that female students outperform male students on writing assessments (e.g., National Center for Education Statistics, 2012). However, the literature offers little evidence regarding differences in the composition processes that might lead to the observed score differences, with the exception of a few studies suggesting an advantage for females in overall writing fluency (e.g., Camarata & Woodcock, 2006). In this study, we analyzed a collection of behavioral features and compared males and females in their writing processes. We approached the question using different methodologies borrowed from data mining and sequence analyses. In the next section, we provide some specific theoretical and research background for this study.

Research Background

Usually about 20–40 minutes are assigned for essay tasks in an assessment. In this timed-writing condition, how effectively writers approach the writing task is of importance. The model in Hayes (2012) specifies three layers of aspects contributing to writing—a *control level*, a *process level*, and a *resource level*. At the process level, Hayes further suggests that, under a certain task environment, a text production process involves four subprocesses: *proposer, translator*, *transcriber,* and *evaluator*. The relationships among these subprocesses are nonlinear but recursive and interleaved. Given that these subprocesses would compete for a limited cognitive resource, for which a scenario can be understandably exacerbated in a timed-writing condition, an effective writing process requires efficient coordination of these subprocesses. Previous research has shown that these subprocesses and their relationships identified in the theoretical model can be operationalized and empirically examined via keystroke logging, in that the time spent on various activities (e.g., planning, editing) is tracked and quantifiable (Zhang & Deane, 2015). Specifically, timing and behavioral features such as the time spent before the start of writing, the median pause length between words or sentence boundaries, the extent of deletion as a function of total number of keystrokes, and frequency of pauses inside of words can be extracted and analyzed under the guidance of a chosen theoretical framework. Examples of keystroke logs are given later (in the "Method" section). The examples show that the log tracks each key press action and cursor movement, as well as associated temporal information, during writing.

Evidence revealed by the timing and process features extracted from keystroke logs can be critical when we seek to understand population group differences in writing performance. In this study, we uncovered and compared patterns in students' writing processes using action sequences extracted from keystroke logging. In particular, we sought to answer the following three research questions:

1. Do features describing students' keystroke actions and timing differ by score or gender?
2. Do features describing students' keystroke actions and timing change over the course of the writing session?
3. What patterns of keystroke action sequences best discriminate among different groups—for example, students of different gender or of different proficiency level indicated by writing scores?

Research on analyzing action sequences has been conducted both in education and other fields. For instance, DiCerbo, Liu, Rutstein, Choi, and Behrens (2011) used digraphs to visualize and analyze sequential process data collected from an assessment. As another example, process management methods and process analysis techniques originated in business management, such as the Petri net, have also been used to study behavioral patterns of learners (Howard, Johnson, & Neitzel, 2010). More generally, in the field of machine learning and data mining, sequential pattern mining techniques have been shown to be effective in identifying frequent patterns from sequence data in log records (Han, Cheng, Xin, & Yan, 2007; Liu, Zhang, Xiong, Jiang, & Yang, 2014). Related methods developed by computer scientists were originally intended to capture the most frequent subsequences or to find the most observed associations among items purchased during online shopping. In this study, we applied sequence analysis to identify patterns in the subsequences of the keystroke sequence data.

Method

Data Collection Instrument

We used a data set collected from a scenario-based assessment (SBA) of writing developed at Educational Testing Service (ETS; Bennett, Deane, & van Rijn, 2016). Among other functions, this particular SBA design intends to model the writing process that expert writers generally follow. Bennett et al. (2016) also provided the theoretical grounding underpinning such an assessment design. Specifically, in the assessment, students are first asked to complete a series of lead-in tasks, all related to a unifying scenario on a familiar topic. The topic in our data set is about choosing the best theme for a school's culture fair event. The lead-in tasks include items that measure component skills of argumentative writing, including the ability to create and evaluate arguments, the ability to critique an argument, and the ability to summarize informational text that contains an argument. The final task in the assessment is to write a persuasive essay on the same topic. The entire assessment is administered in two consecutive 45-minute testing sessions. The data set used in this study is part of a larger data collection in 2013 (van Rijn, Chen, & Yan-Koo, 2016), in which six persuasive and argumentative writing assessments were administered. In this study, we examined one of those assessments, Culture Fair, and focused on the essay task because we were interested only in students' essay-writing processes.

Participants and Data Set

Our data set included a total of 761 sixth- to eighth-grade students from the United States, of whom 42% were female, 39% were male, and 19% did not report their gender. As for race/ethnicity, 56% were White, 18% were Hispanic, 3% were African American, 3% were Asian, fewer than 1% belonged to other groups, and 19% did not indicate their race. In terms of their language proficiency levels, 504 students (66%) were English proficient, 33 (4%) were English language learners (ELLs), 2.5% percent were originally ELL but reclassified to English proficient, and 24% did not report this information. Socioeconomic status (SES) data were available for 576 students, of whom 25% qualified for free or reduced-price lunches. All the demographic information was reported by classroom teachers and collected through a survey.

The completed essays were graded by human raters on two separate dimensions based on a detailed rubric for each dimension. The first rubric score was based on the discourse-level features in a multiparagraph text—grammar and word usage, spelling, syntactic variety, coherence, structure, and so on. The second rubric score was based on genre-specific skills for writing an argumentative/persuasive essay—quality of the reasoning, command of argument structure, and so on. The score scale of both rubrics ranged from 0 to 5. Each rubric was graded by two raters. We randomly selected for analysis one human rater score for each rubric. Further, we excluded all samples with scores of 0 because they were often either blank or very short. We also combined the score categories of 4 and 5 for each rubric due to the small number samples in Score Category 5 in the data set.

Keystroke Logging and Keystroke Actions

Students' writing processes were recorded using ETS's keystroke logging engine. The raw data included time-stamped action logs on all key presses conducted by students. One keystroke record corresponds to one keyboard operation. Examples of keystroke operation include *insert, delete, copy, paste, jump*, and *replace*. The raw keystroke logs included both the actions and the content of these actions—for example, alphanumeric characters being added or removed. Figure 1 provides a simple example of keystroke log records. In this example, the student constructed the sentence "It is an apple."

We can either obtain the composed sentence from the recorded final product or recover the sentence from the keystroke logging record. Besides the final product, the keystroke log reveals the process of how this sentence was produced. From the log, it becomes clear that the student changed the word *the* to the word *an* during writing, for which information would not have been known from the product alone. Figure 2 shows the final sentence and the editing process, as well as related keystroke actions/operations corresponding to the example in Figure 1. The keystroke actions are color coded to ease the reading. Green indicates inserts, and red indicates deletes.

StudentID	Index	TimeStamp	InterKeyInterval	PositionInText	Content	Operation	TextToDate
119	0	0	-	0	I	Insert	I
119	1	0.56	0.56	1	t	Insert	It
119	2	1.55	0.99	2		Insert	It
119	3	2.39	0.84	3	i	Insert	It i
119	4	2.55	0.16	4	s	Insert	It is
119	5	3.22	0.67	5		Insert	It is
119	6	3.39	0.17	6	t	Insert	It is t
119	7	3.56	0.17	7	h	Insert	It is th
119	8	4.28	0.72	8	e	Insert	It is the
119	9	4.72	0.44	7	e	Delete	It is th
119	10	6.45	1.73	6	h	Delete	It is t
119	11	6.72	0.27	5	t	Delete	It is
119	12	9.3	2.58	6	a	Insert	It is a
119	13	10.41	1.11	7	n	Insert	It is an
119	14	10.75	0.34	8		Insert	It is an
119	15	11.19	0.44	9	a	Insert	It is an a
119	16	12.51	1.32	1	p	Insert	It is an ap
119	17	13.08	0.57	11	p	Insert	It is an app
119	18	13.5	0.42	12	l	Insert	It is an appl
119	19	14.09	0.59	13	e	Insert	It is an apple
119	20	14.25	0.16	14	.	Insert	It is an apple.

Figure 1 Example of keystroke logging records.

Final sentence: It is an apple.

Editing process: It is ~~the~~an apple.

Keystroke actions:
Insert-Insert-Insert-Insert-Insert-Insert-Insert-Insert-Insert-Delete-Delete-Delete-Insert-Insert-Insert-Insert-Insert-Insert-Insert-Insert-Insert

Figure 2 Example of keystroke action sequence.

This study focuses on students' writing behaviors and intends to characterize and compare the keystroke actions from different student groups. Because this was an initial study, we purposely ignored the linguistic content of what was written. That is, we did not distinguish an insertion of the space with an alphanumeric character or a punctuation mark. Our analysis took into consideration the keystroke action types, the timing information for these actions, and the sequences of keystroke actions.

Data Analysis

To analyze the keystroke sequence data and answer our research questions, we adopted two approaches. The first approach was to generate variables from the keystroke sequences to characterize a writer's text production processes. For the second approach, we applied sequence mining techniques to directly study the keystroke sequences. Results from the first approach were used to address the first two research questions, and results from the second approach were used to answer the third research question.

Defining the Feature Variables

In the first approach, we generated the following seven variables from each of the keystroke logs. These variables provided a rather general picture of the text production process conducted by a student.

- *Total Time* captures the total active writing time; the active writing time is defined as the time interval between the first and last keystrokes in a log file. Usually, students spent some time before making the first keystroke to prepare for the task. Students may also have been idle after the last keystroke before submitting their final essay. The total active writing time provides an estimate of the effort and persistence level of a student.
- *Num Record* is the total number of keystroke records during the active writing session. Similar to the total writing time, this summary statistic provides a general sense of the writing effort made by the students.
- *Num Insert* and *Num Delete* are the numbers of insertions and deletions among the *Num Record*. Among the various keystroke actions, *insert* and *delete* are the two most common actions. On average, they together accounted for 99% of all the actions across students' keystroke logs. We decided to focus on these two actions.
- *DIRatio* is the ratio of *Num Delete* over *Num Insert*, which approximately captures the extent of editing and revision of any kind.
- *Median IKI* is the median of the inter-key interval, which captures the median lag time between two adjacent keystroke actions.
- *Efficiency* is estimated by the number of keystroke records per second, which indicates a general writing speed. This feature variable and Median IKI are arguably indicators of writing fluency.

To answer our first research question on the differences between subgroups of students with different scores and different genders, we ran ANOVA and ANCOVA analyses to compare the features for students in different subgroups.

The second research question focuses on the sequential patterns of the keystroke actions over time. To answer this research question, we first divided each student's writing session into three subsessions (i.e., beginning, middle, and end), and then generated the feature variables on each subsession. To compare subsessions, we conducted one-way repeated measures ANOVA, in which the dependent variables were the above-mentioned process features and the independent variable was the writing subsession.

Sequence Mining of Keystroke Action Sequences

To address the third question on the group differences in keystroke action sequences, we applied the sequence mining techniques (Han et al., 2007; Uma, Kalaivany, & Aghila, 2013) on the keystroke action sequences. Even though the basic ideas of sequence mining are simple, the implementation can be complex, particularly when the subsequences of interest are long. For short candidate subsequences, enumeration of different subsequences and identification of the most frequent ones in the data set are not as intensive. However, simple and short subsequences do not fully capture the behavioral patterns that we were interested in. In this study, the actions were low-level basic actions. The variety of keystroke action types is limited, but the information embedded in longer subsequences can be very interesting.

With an increase in the length of the subsequences, the difficulty of the problem increases dramatically, which makes it an active research field in data mining. Given certain constraints, such as the maximum length of subsequences, various algorithms (e.g., Uma et al., 2013) have been developed to speed up the analysis processes. In our analysis, we used the R package TraMineR (Gabadinho, Ritschard, Müller, & Studer, 2011), which provides the functions of sequence mining and visualization. As an initial attempt to conduct sequence mining of the keystroke logs, we constrained the maximum subsequence length to 10 actions. We first identified the most frequent subsequences among all keystroke subsequences that were equal to or shorter than 10 actions. Subsequently, analyses were conducted to identify the keystroke subsequences that were most discriminating across different score levels and the two gender groups. The results of discriminant analysis, using the *seqecmpgroup* function in the R package TraMineR, were used to assist in interpretations of the findings and to help understand subgroup differences more completely.

Results

Keystroke Patterns of Students With Different Scores and Genders

As described in the "Method" section under Participants and Data Set, all student essays were scored by human raters on two dimensions based on two rubrics, with Rubric 1 focusing on writing basics and Rubric 2 on content. In our analyses, since we combined score levels 4 and 5, scores on these two rubrics ranged from 1 to 4. The distributions of student scores on both rubrics are shown in Figure 3. For both rubrics, none of the score groups had fewer than 25 students.

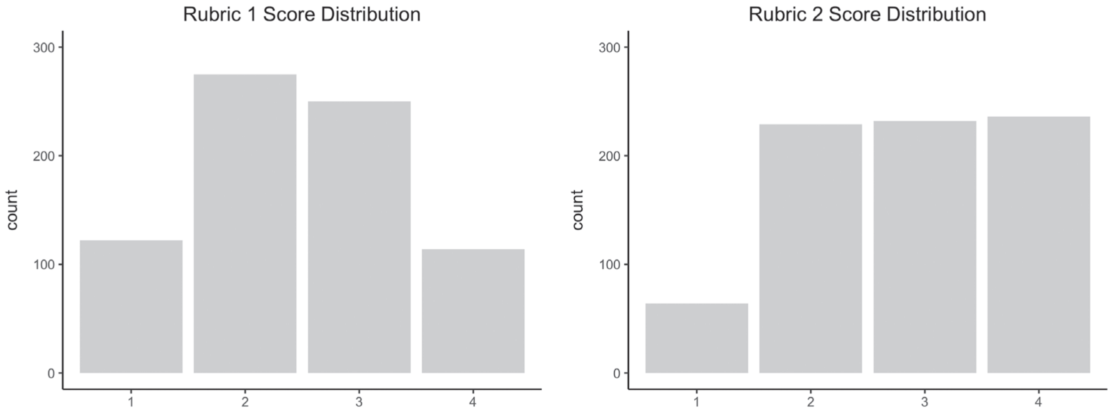


Figure 3 Distributions of rubric scores.

Table 1 Comparisons of Different Score Groups

Feature	Mean	*SD*	Rubric 1	Rubric 2
Total Time	871.98	433.59	$F(3, 757) = 72.91^*$	$F(3, 757) = 45^*$
Num Record	1,237.41	657.72	$F(3, 757) = 9.34^*$	$F(3, 757) = 92.52^*$
Num Insert	1,082.96	558.56	$F(3, 757) = 143.80^*$	$F(3, 757) = 96.73^*$
Num Delete	150.49	128.34	$F(3, 757) = 39.74^*$	$F(3, 757) = 36.67^*$
DIRatio	0.13	0.08	n.s.	n.s.
Median IKI	0.29	0.10	$F(3, 757) = 9.34^*$	$F(3, 757) = 12.84^*$
Efficiency	1.52	0.59	$F(3, 757) = 6.24^*$	$F(3, 757) = 9.07^*$

Note. Total Time = total active writing time; Num Record = total number of keystroke records during Total Time; Num Insert = number of insertions among Num Record; Num Delete = total number of deletions among Num Record; DIRatio = Num Delete over Num Insert; Median IKI = median interkey interval; Efficiency = number of keystroke records per second.
$^*p < .001$.

To compare the keystroke patterns of students with different scores, we ran a series of ANOVAs with the seven keystroke feature variables as dependent variables and the scores as independent variables. The mean, standard deviation, and ANOVA results for all feature variables on both rubrics are summarized in Table 1.

We found that six out of seven keystroke feature variables were significantly different among score groups with $p < .001$: namely Total Time, Num Record, Num Insert, Num Delete, Median IKI, and Efficiency. Even after the Bonferroni correction of the seven sets of comparisons on the same data sets (α level needs to be adjusted to α/m, where m is the number of tests), the results were still significant at the $.05/7 = .007$ level for the significant findings. The difference on DIRatio was not significant, with $p > .007$ on either rubric. Given the statistically significant ANOVA test results, post hoc analyses were conducted using the Tukey HSD tests on all pairwise comparisons. The post hoc Tukey HSD test results showed that, for both rubrics, higher scoring students spent more time on active writing (Total Time) compared to the lower scoring students, and the results differed significantly for all pairwise comparisons at the .05 level. For the records of keystrokes, higher scoring students had greater numbers of keystroke records (Num Record), more insert records (Num Insert), and more delete records (Num Delete). The results were also significant for all pairwise comparisons at the .05 level. In terms of writing fluency, generally, high-performing students had shorter interkey gaps (Median IKI) and typed more per second (Efficiency). In taking a closer look, the results showed that not all pairwise comparisons were significant.

On the results of Median IKI (as shown in Figure 4, error bars indicating plus and minus one standard error), for scores using Rubric 1, the differences between Score Groups 1 and 2 and between Score Groups 3 and 4 were not significant at the .05 level. For scores using Rubric 2, the differences between Score Groups 1 and 2 and between Score Groups 2 and 3 were not significant at the .05 level. Consistently over both rubrics, Score Group 4 had shorter Median IKI than both Score Groups 1 and 2, and Score Group 3 had shorter Median IKI than Score Group 1.

On the results of Efficiency (as shown in Figure 5, error bars indicating plus and minus one standard error), for scores using Rubric 1, the significant differences at the .05 level were between Score Groups 1 and 4, Score Groups 2 and 4, and

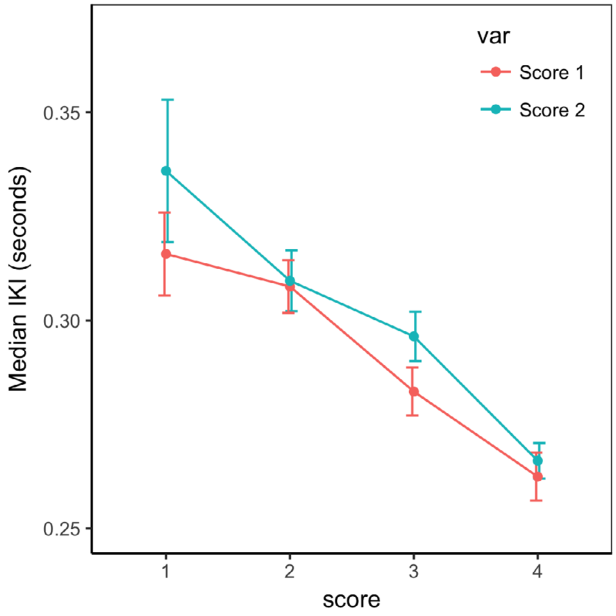


Figure 4 Median interkey interval conditional on score level.

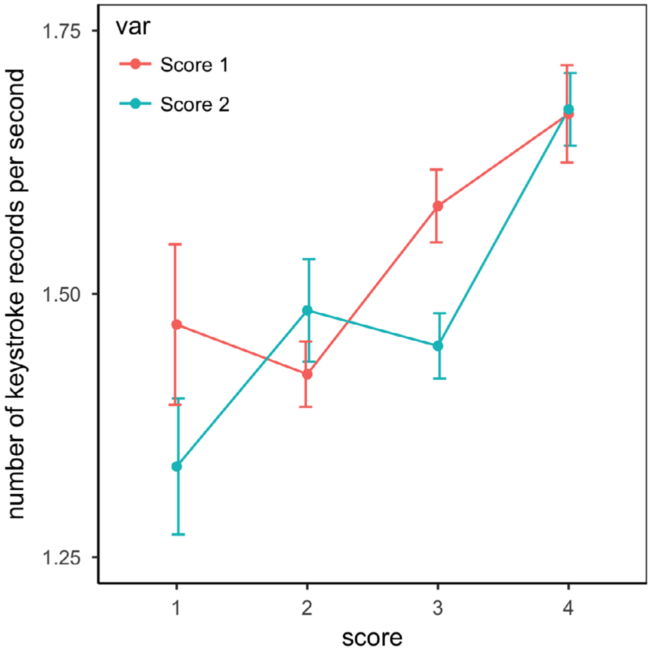


Figure 5 Efficiency conditional on score level.

Score Groups 2 and 3. For scores using Rubric 2, the significant differences at the .05 level were between Score Groups 1 and 4, Score Groups 2 and 4, and Score Groups 3 and 4. Consistently over both rubrics, Score Group 4 had greater writing efficiency (i.e., more keystroke per second) than both Score Groups 1 and 2.

As for the male versus female comparisons, two sample independent t-tests were conducted to compare the group means. As summarized in Table 2, the results showed that, compared to male students, female students spent longer times on writing and had greater numbers of insertions, deletions, and revisions/edits. Female students also typed faster than male students with shorter median between-key gaps and great number of keystroke records per second. However, using $\alpha = .007$ level with the Bonferroni correction, the differences of Total Time were no longer significant.

Table 2 Comparisons of Different Gender Groups

Feature	Female mean	Female *SD*	Male mean	Male *SD*	*t*	*df*	*p*
Total Time	951.98	426.46	863.94	430.09	2.55	611	0.01
Num Record	1,442.77	721.01	1,121.89	572.95	6.13	594	<0.001
Num Insert	1,248.02	608.58	994.53	494.93	5.68	598	<0.001
Num Delete	189.56	148.10	124.02	105.89	6.33	569	<0.001
DIRatio	0.15	0.08	0.12	0.08	4.54	611	<0.001
Median IKI	0.28	0.08	0.31	0.12	4.42	511	<0.001
Efficiency	1.59	0.55	1.42	0.62	3.70	596	<0.001

Note. Total Time = total active writing time; Num Record = total number of keystroke records during Total Time; Num Insert = number of insertions among Num Record; Num Delete = total number of deletions among Num Record; DIRatio = Num Delete over Num Insert; Median IKI = median interkey interval; Efficiency = number of keystroke records per second.

We also found that female students received significantly higher scores on both rubrics than male students. Chi-square tests on the relations between gender and scores revealed that, for both scoring rubrics, the relations were significant at the .05 level, χ^2 $(3, N = 761) = 11.32$, $p = .01$ for Rubric 1, and χ^2 $(3, N = 761) = 22.42$, $p < .01$ for Rubric 2. Hence, it is of interest whether the observed differences in writing process features, described above, are a function of score differences. To address this question, we conducted a series of ANCOVAs while controlling for rubric scores. In the ANCOVAs, feature variables were treated as dependent variables, scores as covariate, and gender as independent variable.

Most of the results, as shown in Table 2, remain significant at the .001 level, except for Total Time, $F(1, 610) = 1.53$, $p = .22$ with Score 1 as covariate, and $F(1, 610) = 1.24$, $p = .27$ with Score 2 as covariate. The gender differences among the remaining keystroke features were significant at the .001 level after controlling the score differences: for Num Record, $F(1, 610) = 26.85$, $p < .001$ with Score 1 as covariate, and $F(1, 610) = 21.32$, $p < .001$ with Score 2 as covariate; for Num Insert, $F(1, 610) = 20.82$, $p < .001$ with Score 1 as covariate, and $F(1, 610) = 16.82$, $p < .001$ with Score 2 as covariate; for Num Delete, $F(1, 610) = 29.65$, $p < .001$ with Score 1 as covariate, and $F(1, 610) = 26.86$, $p < .001$ with Score 2 as covariate; for DIRatio, $F(1, 610) = 20.0$, $p < .001$ with Score 1 as covariate, and $F(1, 610) = 18.54$, $p < .001$ with Score 2 as covariate; for Median IKI, $F(1, 610) = 15.39$, $p < .001$ with Score 1 as covariate, and $F(1, 610) = 14.57$, $p < .001$ with Score 2 as covariate; for Efficiency, $F(1, 610) = 10.79$, $p < .001$ with Score 1 as covariate, and $F(1, 610) = 10.89$, $p < .001$ with Score 2 as covariate. The above results held significant using the α level of .007 with the Bonferroni correction.

Changes of Keystroke Features Over Time

Besides the overall features of the keystroke action patterns, we were also interested in whether, and how, students may change behaviors over the course of writing. Therefore, we divided each keystroke log evenly into three subsessions (with an equal length in time) using the Total Time representing the beginning, middle, and end of a writing session and calculated the feature values separately for each of the three subsessions. Because the features were computed on three subsessions of the same students and were related to each other, we used repeated measures ANOVAs to test relations between subsessions and keystroke features.

The repeated measures ANOVA results are summarized in Table 3. Because we divided the writing session based on time, the variable Total Time was no longer relevant and hence was excluded from this analysis. When looking at writing behavior across time, the significant differences at the .001 level were found on the number of records and the number of inserts across the beginning, middle, and end of the writing subsessions.

Posterior Tukey tests on the significant findings showed that the differences for Num Record and Num Insert were between the last and the first two subsessions, with more total records and inserts in the last subsession (as shown in Figure 6). For Efficiency, however, the differences were between the middle subsession and the other two subsessions, with significantly fewer keystrokes typed per second in the middle subsession (as shown in Figure 7).

Sequence Mining of Keystroke Actions

Besides statistical analysis on the generated keystroke feature variables, we further analyzed the keystroke sequences by identifying patterns in the subsequences of the keystroke activities. We considered the most frequent subsequences

Table 3 Results of the Differences Among Subsessions Using Repeated Measures ANOVA

Feature	χ^2	*df*	*p*
Num Record	24.15	2	<0.001
Num Insert	31.37	2	<0.001
Num Delete	5.97	2	0.05
DIRatio	1.87	2	0.39
Median IKI	1.99	2	0.37
Efficiency	8.84	2	0.01

Note. Total Time = total active writing time; Num Record = total number of keystroke records during Total Time; Num Insert = number of insertions among Num Record; Num Delete = total number of deletions among Num Record; DIRatio = Num Delete over Num Insert; Median IKI = median interkey interval; Efficiency = number of keystroke records per second.

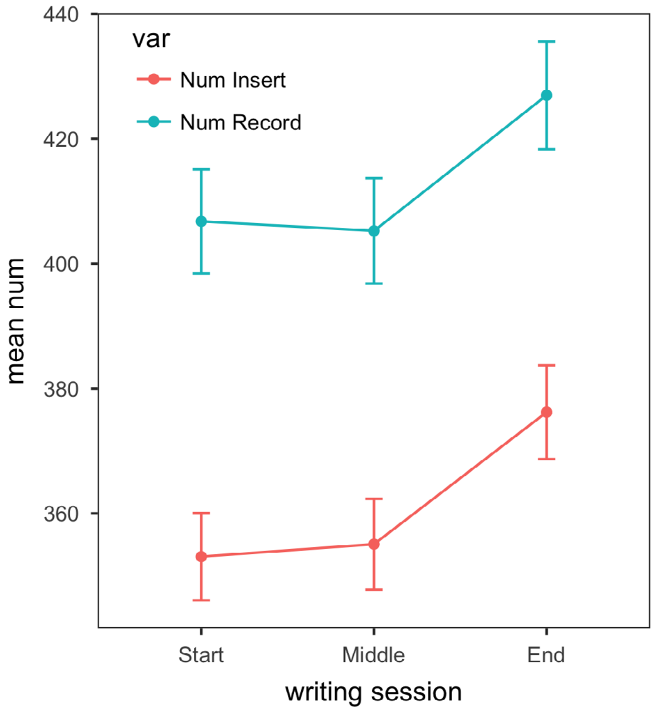


Figure 6 Subsession differences for Num Insert and Num Record. Error bars represent the standard errors.

and the subsequences with the most discriminating power. In the following analysis, we only consider the actions, while ignoring the lengths of the pauses between actions. Figure 8 shows several examples illustrating the full keystroke action sequences of the essays. Each sequence represents all of one student's keystroke actions in completing the assigned essay. Each student's action sequence consists of one or more of the five keystroke actions—that is, insert, delete, cut, paste, and replace. The action types are color coded. The most frequent action in the sequences was insert, followed by delete. The other three action types were rather rare.

To do sequence mining, two parameters are usually required to make sure that the calculations are manageable. One is the *minimum support threshold*, in which support is calculated as the percentage of sequences containing the subsequence. The other is the *maximum subsequence length*. Searching for the subsequences becomes more complex, with lower minimum support thresholds and larger maximum subsequence lengths, because the program needs to search for a larger variety of subsequences. The subsequence search and calculating process can become quite time consuming as well. However, with higher minimum support thresholds and shorter maximum subsequence lengths, meaningful patterns might be left undiscovered. In this study, we used a rather low minimum support threshold and a relatively large maximum subsequence length (about twice the average word length in this data set). Given the size of the candidate sequences, the

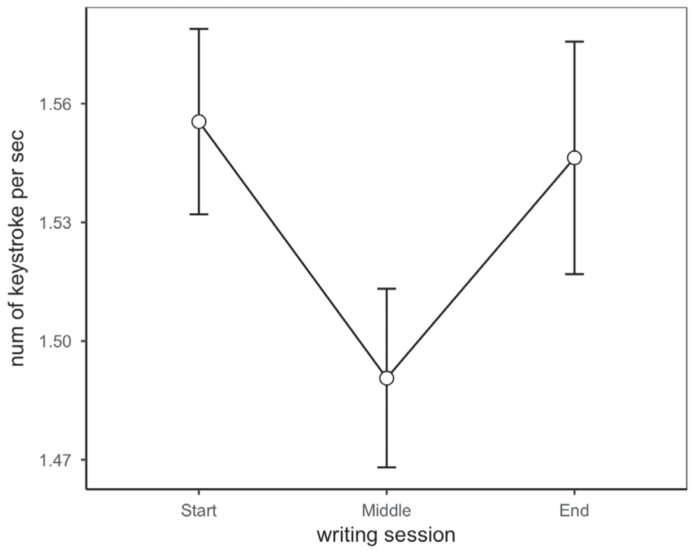


Figure 7 Subsession differences for number of keystrokes per second. Error bars represent the standard errors.

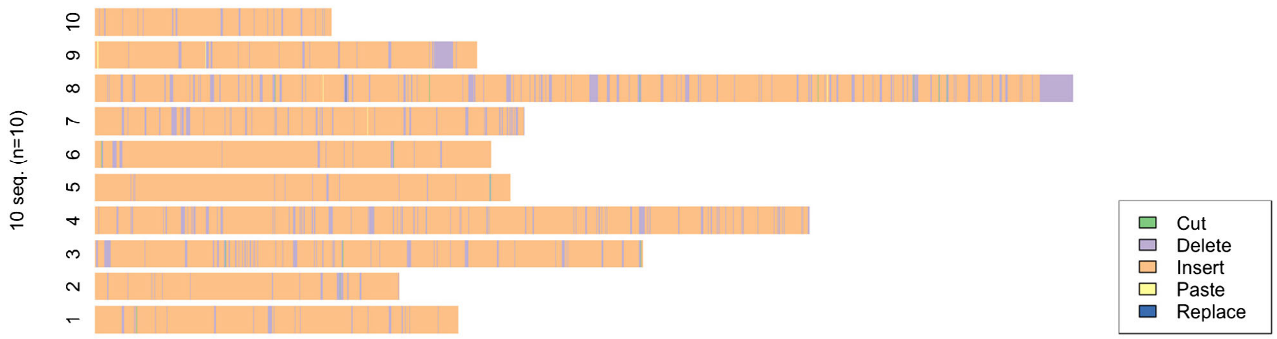
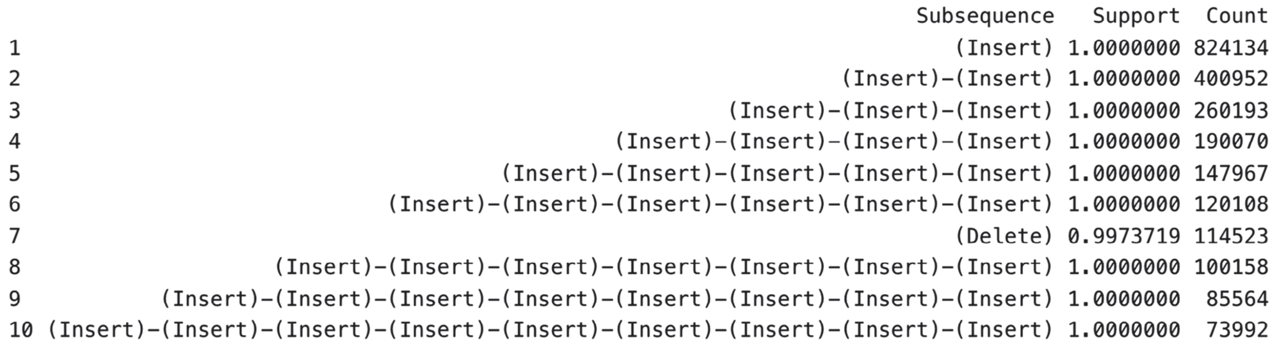


Figure 8 Keystroke action sequences for a random sample of 10 essays as examples.



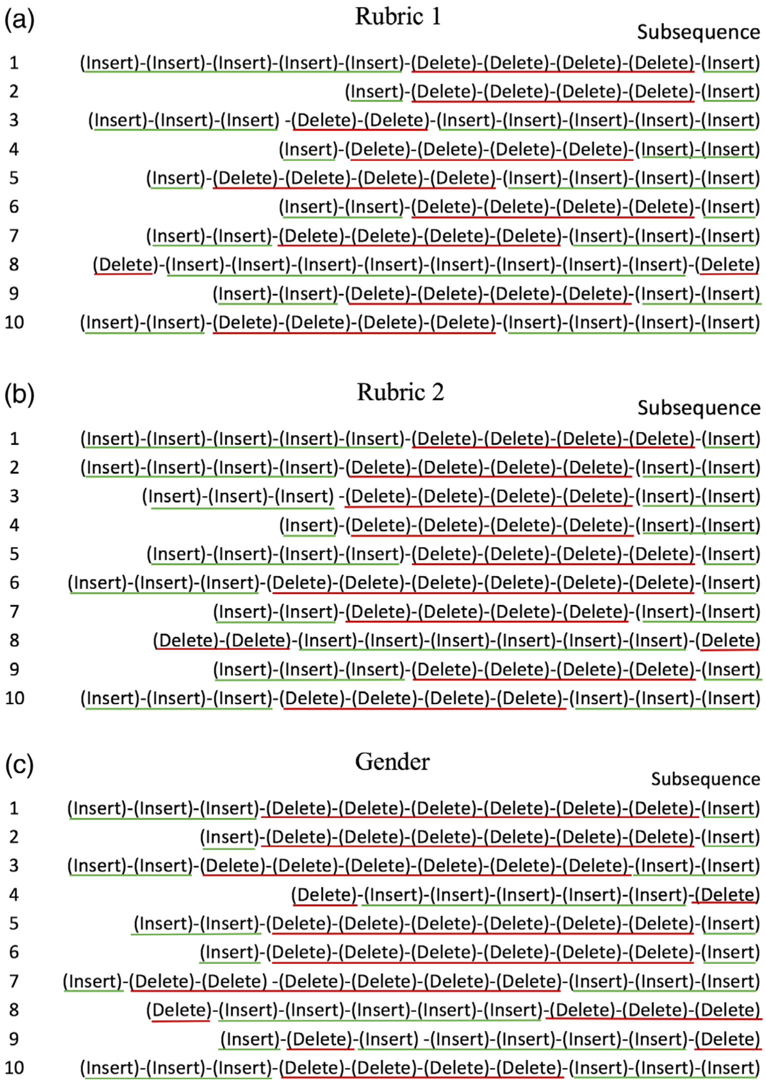
```
                                                                  Subsequence   Support  Count
1                                                                    (Insert) 1.0000000 824134
2                                                           (Insert)-(Insert) 1.0000000 400952
3                                                  (Insert)-(Insert)-(Insert) 1.0000000 260193
4                                         (Insert)-(Insert)-(Insert)-(Insert) 1.0000000 190070
5                                (Insert)-(Insert)-(Insert)-(Insert)-(Insert) 1.0000000 147967
6                       (Insert)-(Insert)-(Insert)-(Insert)-(Insert)-(Insert) 1.0000000 120108
7                                                                    (Delete) 0.9973719 114523
8              (Insert)-(Insert)-(Insert)-(Insert)-(Insert)-(Insert)-(Insert) 1.0000000 100158
9     (Insert)-(Insert)-(Insert)-(Insert)-(Insert)-(Insert)-(Insert)-(Insert) 1.0000000  85564
10 (Insert)-(Insert)-(Insert)-(Insert)-(Insert)-(Insert)-(Insert)-(Insert)-(Insert) 1.0000000  73992
```

Figure 9 Top 10 most frequent subsequences.

time spent on the subsequence searching and calculation based on the set parameters were acceptable. Further raising the minimum subsequence length to beyond 10 would substantially increase the programming time.

We first calculated the most frequent subsequences for all keystroke sequences, with the minimum support threshold as 0.1 and the maximum subsequence length as 10. The results are shown in Figure 9.

Nine out of the 10 most frequent subsequences were straight insert actions; the only exception was a single delete action. The subsequences of inserts appeared in all sequences, and delete actions appeared in 99.74% of all sequences. These findings on the most frequent subsequences were intuitive and less interesting.



(a) Rubric 1

	Subsequence
1	(Insert)-(Insert)-(Insert)-(Insert)-(Insert)-(Delete)-(Delete)-(Delete)-(Delete)-(Insert)
2	(Insert)-(Delete)-(Delete)-(Delete)-(Delete)-(Insert)
3	(Insert)-(Insert)-(Insert) -(Delete)-(Delete)-(Insert)-(Insert)-(Insert)-(Insert)-(Insert)
4	(Insert)-(Delete)-(Delete)-(Delete)-(Delete)-(Insert)-(Insert)
5	(Insert)-(Delete)-(Delete)-(Delete)-(Delete)-(Insert)-(Insert)-(Insert)-(Insert)
6	(Insert)-(Insert)-(Delete)-(Delete)-(Delete)-(Delete)-(Insert)
7	(Insert)-(Insert)-(Delete)-(Delete)-(Delete)-(Delete)-(Insert)-(Insert)-(Insert)
8	(Delete)-(Insert)-(Insert)-(Insert)-(Insert)-(Insert)-(Insert)-(Insert)-(Insert)-(Delete)
9	(Insert)-(Insert)-(Delete)-(Delete)-(Delete)-(Delete)-(Insert)-(Insert)
10	(Insert)-(Insert)-(Delete)-(Delete)-(Delete)-(Delete)-(Insert)-(Insert)-(Insert)-(Insert)

(b) Rubric 2

	Subsequence
1	(Insert)-(Insert)-(Insert)-(Insert)-(Insert)-(Delete)-(Delete)-(Delete)-(Delete)-(Insert)
2	(Insert)-(Insert)-(Insert)-(Insert)-(Delete)-(Delete)-(Delete)-(Delete)-(Insert)-(Insert)
3	(Insert)-(Insert)-(Insert) -(Delete)-(Delete)-(Delete)-(Delete)-(Insert)-(Insert)
4	(Insert)-(Delete)-(Delete)-(Delete)-(Delete)-(Insert)-(Insert)
5	(Insert)-(Insert)-(Insert)-(Insert)-(Delete)-(Delete)-(Delete)-(Delete)-(Insert)
6	(Insert)-(Insert)-(Insert)-(Delete)-(Delete)-(Delete)-(Delete)-(Delete)-(Delete)-(Insert)
7	(Insert)-(Insert)-(Delete)-(Delete)-(Delete)-(Delete)-(Insert)-(Insert)
8	(Delete)-(Delete)-(Insert)-(Insert)-(Insert)-(Insert)-(Insert)-(Insert)-(Delete)
9	(Insert)-(Insert)-(Insert)-(Delete)-(Delete)-(Delete)-(Delete)-(Insert)
10	(Insert)-(Insert)-(Insert)-(Delete)-(Delete)-(Delete)-(Delete)-(Insert)-(Insert)-(Insert)

(c) Gender

	Subsequence
1	(Insert)-(Insert)-(Insert)-(Delete)-(Delete)-(Delete)-(Delete)-(Delete)-(Delete)-(Insert)
2	(Insert)-(Delete)-(Delete)-(Delete)-(Delete)-(Delete)-(Delete)-(Insert)
3	(Insert)-(Insert)-(Delete)-(Delete)-(Delete)-(Delete)-(Delete)-(Delete)-(Insert)-(Insert)
4	(Delete)-(Insert)-(Insert)-(Insert)-(Insert)-(Insert)-(Delete)
5	(Insert)-(Insert)-(Delete)-(Delete)-(Delete)-(Delete)-(Delete)-(Delete)-(Insert)
6	(Insert)-(Delete)-(Delete)-(Delete)-(Delete)-(Delete)-(Delete)-(Insert)
7	(Insert)-(Delete)-(Delete) -(Delete)-(Delete)-(Delete)-(Delete)-(Insert)-(Insert)-(Insert)
8	(Delete)-(Insert)-(Insert)-(Insert)-(Insert)-(Insert)-(Delete)-(Delete)-(Delete)
9	(Insert)-(Delete)-(Insert) -(Insert)-(Insert)-(Insert)-(Insert)-(Delete)
10	(Insert)-(Insert)-(Insert)-(Delete)-(Delete)-(Delete)-(Delete)-(Insert)-(Insert)-(Insert)

Figure 10 Top 10 most discriminant subsequences.

Next, we ran three separate analyses to find the subsequences with the most discriminating power for students on different score levels and between the two gender groups. The discrimination power of a certain subsequence was tested with the Pearson Chi-square of the table that cross-tabulates the presence and absence of the subsequence. Lower p values indicated more discriminating power for the subsequence. The identified subsequences with the most discriminating power for the scores on Rubric 1, Rubric 2, and the two gender groups are shown in Figure 10. The insert and delete actions are, respectively, underscored by green and red ink.

The results indicate that the most discriminating subsequences featured longer sequences (more than an average word length of 4.5 characters in this data set) of delete or insert actions rather than local editing featuring alternating deletes and inserts. In addition, the supports on all the most discriminating subsequences were high. They were near or well above 0.5, which means that these subsequences also appeared in more than half of the sequences and were not rare. The p values of the discriminating tests were also <.001.

Discussion

In this study, we made use of keystroke logging data and examined students' essay-writing processes in an SBA. Comparisons were made between groups of different proficiency levels and between the two gender groups. This study serves as one step further toward addressing a larger research question on closing achievement gaps of different population groups, as findings of this study have implications for gaining deeper understanding of observed group differences and for designing interventions to close the gaps among population groups.

Built upon previous research findings, we addressed three research questions. For the first question, we extracted seven feature variables from the keystroke logs. Compared to students in lower scoring groups and male students, students in higher scoring groups and female students tended to spend longer times on writing and use their time more efficiently, while exhibiting greater writing fluency. This result is largely consistent with previous findings, with the exception that current analyses showed no statistical difference in the ratio of deletion and insertion records among score groups, only between the two gender groups. Being considered as a proxy for the extent of editing behavior, one possible explanation is that this statistic is too general, therefore hiding nuanced differences between score groups that can only be detectable using more specific metrics measuring editing.

In the second question, each student's writing process was evenly divided into three stages — start, middle, and end — based on the total active writing time. Results revealed significantly greater absolute numbers of keystroke actions (insertion and deletion) made by students toward the end of their writing session, with the greatest number of keystrokes per second in the middle of their writing process. Students appeared to be most efficient in text production after some time entering the writing task (in the middle of the writing course) and most "busy" with generating text right before submission (toward the end), although not necessarily in the most efficient way.

The third research question tackled the keystroke action sequences. The most striking discovery was that the sequence patterns that were most discriminating across score and gender groups were word-length deletion sequences bounded by insertions. This result suggests that potentially the most distinguishing features between high and low score responses are not minor edits, which tend to be quick fixes to one or two characters. A future research direction will be to analyze the linguistic context of these deletion activities. One hypothesis is that these long deletion sequences are followed by long jumps, because the writing research has suggested that expert and novice writers differ in their editing and revision processes. Novice writers tend to conduct mainly small and local edits, whereas expert writers pay more attention to the global editing throughout the texts, signaled by jumping from one place to another to make content modifications (Breetvelt, van den Bergh, & Rijlaarsdam, 1994).

In the current analyses, when constructing the keystroke action feature variables, we intentionally ignored the contents of the keystroke actions. For future studies, it would be interesting to introduce linguistic analysis to study not only how students edited but also what they edited. The keystroke actions, such as insert and delete, can then be associated with their functions. For instance, it can be revising typographic errors, or revisions of a whole sentence in terms of the content. For sequence analysis, the current method considered the keystroke actions as discrete events with equal intervals. In future studies, we plan to integrate timing data between adjacent events into sequence mining analysis. By including more information into the future analyses, we hope to deepen our understanding of students' writing behaviors and strategies. With this understanding, we can start to develop appropriate interventions by providing suggestions such as better time management and writing strategies to the low-performing students.

Acknowledgments

We thank our colleagues Hongwen Guo and Randy Bennett for their insightful suggestions.

References

Allen, L. K., Mills, C., Jacovina, M. E., Crossley, S., D'Mello, S., & McNamara, D. S. (2016). Investigating boredom and engagement during writing using multiple sources of information: The essay, the writer, and keystrokes. In Proceedings of the Sixth International Conference on Learning Analytics & Knowledge (pp. 114 – 123). https://doi.org/10.1145/2883851.2883939

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

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 processes. *Journal of Educational Psychology, 103*, 415 – 428. https://doi.org/10.1037/a0022545

Bennett, R. E., Deane, P., & W. van Rijn, P. (2016). From cognitive-domain theory to assessment practice. *Educational Psychologist, 51*, 82 – 107. https://doi.org/10.1080/00461520.2016.1141683

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. https://doi.org/10.1207/s1532690xci1202_2

Camarata, S., & Woodcock, R. (2006). Sex differences in processing speed: Developmental effects in males and females. *Intelligence, 34*, 231 – 252. https://doi.org/10.1016/j.intell.2005.12.001

Chukharev-Hudilainen, E. (2014). Pauses in spontaneous written communication: A keystroke logging study. *Journal of Writing Research, 6*, 61 – 84. https://doi.org/10.17239/jowr-2014.06.01.3

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. https://doi.org/10.1002/ets2.12071

DiCerbo, K. E., Liu, J., Rutstein, D. W., Choi, Y., & Behrens, J. T. (2011, April). *Visual analysis of sequential log data from complex performance assessments*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Dragsted, B., & Carl, M. (2013). Towards a classification of translation styles based on eye-tracking and keylogging data. *Journal of Writing Research, 5*, 133 – 158. https://doi.org/10.17239/jowr- 2013.05.01.6

Ercikan, K., & Pellegrino, J. W. (Eds.). (2017). *Validation of score meaning for the next generation of assessments. The use of response processes*. New York, NY: Routledge. https://doi.org/10.4324/9781315708591

Gabadinho, A., Ritschard, G., Müller, N. S., & Studer, M. (2011). Analyzing and visualizing state sequences in R with TraMineR. *Journal of Statistical Software, 40*(4), 1 – 37. https://doi.org/10.18637/jss.v040.i04

Grabowski, J. (2008). The internal structure of university students' keyboard skills. *Journal of Writing Research, 1*, 27 – 52. https://doi.org/10.17239/jowr-2008.01.01.2

Guo, H., Deane, P., van Rijn, P. W., Zhang, M., & Bennett, R. E. (2018). Modeling basic writing processes from keystroke logs. *Journal of Educational Measurement, 55*, 194 – 216. https://doi.org/10.1111/jedm.12172

Guo, H., Rios, J. A., Haberman, S. J., Liu, O. L., Wang, J., & Paek, I. (2016). A new procedure for detection of students' rapid guessing responses using response time. *Applied Measurement in Education, 29*, 173 – 183. https://doi.org/10.1080/08957347.2016.1171766

Han, J., Cheng, H., Xin, D., & Yan, X. (2007). Frequent pattern mining: Current status and future directions. *Data Mining and Knowledge Discovery, 15*, 55 – 86. https://doi.org/10.1007/s10618-006-0059-1

Hao, J., Smith, L., Mislevy, R., von Davier, A., & Bauer, M. (2016). *Taming log files from game/simulation-based assessments: Data models and data analysis tools* (Research Report No. RR – 16-10). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12096

Hayes, J. R. (2012). Modeling and remodeling writing. *Written Communication*, *29*, 369 – 388. https://doi.org/10.1177/0741088312451260

Howard, L., Johnson, J., & Neitzel, C. (2010). Examining learner control in a structured inquiry cycle using process mining. In *Proceedings of the 3rd International Conference on Educational Data Mining* (pp. 71 – 80). Retrieved from http://educationaldatamining.org/EDM2010/uploads/proc/edm2010_submission_28.pdf

Lee, Y.-H., & Haberman, S. J. (2016). Investigating test-taking behaviors using timing and process data. *International Journal of Testing, 16*, 240 – 267. https://doi.org/10.1080/15305058.2015.1085385

Leighton, J. P., Tang, W., & Guo, Q. (2017). Response processes and validity evidence: Controlling for emotions in think aloud interviews. In B. D. Zumbo & A. M. Hubley (Eds.), *Understanding and investigating response processes in validation research* (pp. 137 – 157). New York, NY: Springer. https://doi.org/10.1007/978-3-319-56129-5_8

Leijten, M., & Van Waes, L. (2013). Keystroke logging in writing research: Using Inputlog to analyze and visualize writing processes. *Written Communication, 30*, 358 – 392. https://doi.org/10.1177/0741088313491692

Leijten, M., Van Waes, L., Schriver, K., & Hayes, J. R. (2013). Writing in the workplace: Constructing documents using multiple digital sources. *Journal of Writing Research, 5*, 285 – 337. https://doi.org/10.17239/jowr-2014.05.03.3

Liu, C., Zhang, K., Xiong, H., Jiang, G., & Yang, Q. (2014). Temporal skeletonization on sequential data: Patterns, categorization, and visualization. *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Min*ing, 28, 1336 – 1345. https://doi.org/10.1145/2623330.2623741

Miller, K. S. (2000). Academic writers on-line: Investigating pausing in the production of text. *Language Teaching Research, 4*, 123 – 148. https://doi.org/10.1177/136216880000400203

National Center for Education Statistics. (2012). *The nation's report card: Writing 2011*. Washington, DC: Author. https://doi.org/NCES2008-468

Ranger, J., Kuhn, J. T., & Gaviria, J. L. (2015). A race model for responses and response times in tests. *Psychometrika, 80*, 791 – 810. https://doi.org/10.1007/s11336-014-9427-8

Roca de Larios, J., Manchón, R., Murphy, L., & Marín, J. (2008). The foreign language writer's strategic behaviour in the allocation of time to writing processes. *Journal of Second Language Writing, 17*, 30 – 47. https://doi.org/10.1016/j.jslw.2007.08.005

Uma, V., Kalaivany, M., & Aghila, G. (2013). Survey of sequential pattern mining algorithms and an extension to time interval based mining algorithm. *International Journal of Advanced Research in Computer Science and Software Engineering, 3*, 1178 – 1183.

Van der Maas, H. L. J., Molenaar, D., Maris, G., Kievit, R. A., & Borsboom, D. (2011). Cognitive psychology meets psychometric theory: On the relation between process models for decision making and latent variable models for individual differences. *Psychological Review, 118*, 339–356. https://doi.org/10.1037/a0022749

Van Rijn, P., Chen, J., & Yan-Koo, Y. (2016). *Statistical results from the 2013 CBAL® English language arts multistate study: Parallel forms for policy recommendation writing* (Research Memorandum No. RM-16-01). Princeton, NJ: Educational Testing Service.

Van Waes, L., Leijten, M., & Van Weijen, D. (2009). Keystroke logging in writing research: Observing writing process with Inputlog. *German as a Foreign Language, 2*(3), 41–64.

Xu, C., & Ding, Y. (2014). An exploratory study of pauses in computer-assisted EFL writing. *Language, Learning and Technology, 18*(3), 80–96.

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. https://doi.org/10.1002/ets2.12075

Zumbo, B. D., & Hubley, A. M. (2017). *Understanding and investigating response processes in validation research*. New York, NY: Springer. https://doi.org/10.1007/978-3-319-56129-5

Suggested citation:

Zhu, M., Zhang, M., & Deane, P. (2019). *Analysis of keystroke sequences in writing logs* (Research Report No. RR-19-11). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12247

Action Editor: James Carlson

Reviewers: Jiangang Hao and Diego Zapata Rivera