Using Keystroke Analytics to Understand Cognitive Processes during Writing

Mo Zhang, Hongwen Guo, and Xiang Liu
Educational Testing Service (Princeton, NJ, 08541)
MZhang, HGuo, XLiu003@ETS.org

ABSTRACT

We present an empirical study on the use of keystroke analytics to capture and understand how writers manage their time and make inferences on how they allocate their cognitive resources during essay writing. The results suggest three distinct longitudinal patterns of writing process that describe how writers approach an essay task in a writing assessment. Discussion of the potential applications of keystroke analytics for improving teaching, learning, and assessing writing are also provided.

Keywords

keystroke logging, writing, cognitive process, time management, writing pattern

1. INTRODUCTION

The study of writing process has long been of interest to the writing research community (e.g., [4], [18], [12], [19], [22]). With the advances in technology, keystroke logging has become a practical and popular tool to capture and study the process of composition in a wide range of contexts [10]. In this study, we demonstrate some research findings on the use of keystroke analytics to understand writers' time management during their writing process. The results have practical implications for the teaching and learning of writing in classrooms.

Previous research on writing cognition suggested several subprocesses of writing [9], including task analysis, text planning, idea generation, translating ideas into natural language, transcribing langauge onto paper (handwriting) or a screen (keyboard-based writing), text revision, copy editing and reviewing. Figure 1 illustrates a simplified version of Hayes coginitive writing model, which specified four main subprocesses of writing. Specifically, idea generation and task preparation (i.e., *proposer*) often manifest as pauses at the start of writing and at sentence boundaries; fluency of putting ideas into language (i.e., *translator*) primarily relates to the size of long sequences of text production without major interruption (also known as "burst"); orthographic proficiency and motor skill (i.e., *transcriber*) typically relates to pauses inside a word and to edits designed to make immediate corrections to spelling errors or typos; and editing and reviewing (i.e., *evaluator*) usually show up as jumps to different locations in the text to make changes or replace large chunks of existing text with new content.

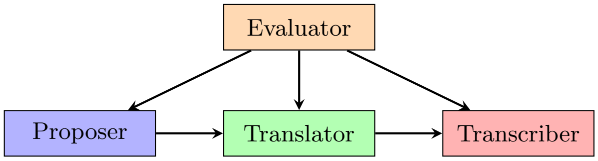

Figure 1: A Cognitive Model of Writing Process

One important implication from the cognitive model is that writing is not a linear process and successful writing calls for effecitive management and coordiation of the subprocesses. Drawing from the cognitive theories of writing, an overview of the types of activities occuring during text composition can be found in [5]. The cognitive resources, as stated in [3], required to carry out each activity do not distribute randomly over the text-production process. Writers often need to decide on which goals to prioritize at which time point because they simply do not have unlimited working memory to accomplish everything at once [11]. With the availability of keystroke logs, how writers distribute their time and cognitive resource to various subprocesses of writing can be quantified and analyzed, which is described in the next section. In this study, we aim to tackle a specific *research question* of whether there are distinct writing-process patterns that may be detected with regard to how writers allocate their cognitive resource to various subprocesses of writing. An identification of meaningful writing profiles will have practical implications for instructors to design curriculum suitable to their classes, and personalize their instruction for learners with different needs and characteristics.

1.1 Keystroke Logging

We consider keystroke logging as a recording of every keypress that the writer makes on the keyboard. The gap time between two consecutive keypresses is often called an interkey interval (IKI), which is also recorded in keystroke logging. A single keystroke record in JSON format may look like this: {"p": "1", "o": "", "n": "I", "t": "0.57"}, where

Mo Zhang, Hongwen Guo and Xiang Liu "Using Keystroke Analytics to Understand Cognitive Processes during Writing". 2021. In: Proceedings of The 14th International Conference on Educational Data Mining (EDM21). International Educational Data Mining Society, 384-390. https://educationaldatamining.org/edm2021/
EDM '21 June 29 - July 02 2021, Paris, France

Table 1: An Example Keystroke Log Segment

Index	PosInText	Content	ContentLen	TimeStamp	ActionType	Context	WordIntended	CursorJump	TextToDate
0	1	I	1	0.57	Insert	WordStart	It	N	I
1	2	t	1	0.62	Insert	InWord	It	N	It
2	3	(*space*)	1	0.99	Insert	BetweenWords		N	It
3	4	i	1	1.30	Insert	WordStart	is	N	It i
4	5	s	1	1.42	Insert	InWord	is	N	It is
5	6	(*space*)	1	1.61	Insert	BetweenWords		N	It is
6	7	a	1	2.12	Insert	WordStart	a	N	It is a
7	8	(*space*)	1	2.22	Insert	BetweenWords		N	It is a
8	9	g	1	2.35	Insert	WordStart	fun	N	It is a g
9	10	o	1	2.43	Insert	InWord	fun	N	It is a go
10	11	o	1	2.55	Insert	InWord	fun	N	It is a goo
11	12	d	1	2.68	Insert	InWord	fun	N	It is a good
12	12	d	1	3.01	Delete	InWord	fun	N	It is a goo
13	11	o	1	3.16	Delete	InWord	fun	N	It is a go
14	10	o	1	3.30	Delete	InWord	fun	N	It is a g
15	9	g	1	3.53	Delete	InWord	fun	N	It is a
16	10	f	1	3.71	Insert	InWord	fun	N	It is a f
17	11	u	1	3.93	Insert	InWord	fun	N	It is a fu
18	12	n	1	4.11	Insert	InWord	fun	N	It is a fun
19	13	(*space*)	1	4.30	Insert	BetweenWords		N	It is a fun
20	14	d	1	4.49	Insert	WordStart	day	N	It is a fun d
21	15	a	1	4.62	Insert	InWord	day	N	It is a fun da
22	16	y	1	4.90	Insert	InWord	day	N	It is a fun day
23	17	.	1	5.13	Insert	PuncMark		N	It is a fun day.

Note: ContentLen=Content Length. ContentLen usually takes the value of 1 unless the writer cuts or pastes in a chunk of text with more than one character. PuncMark=Punctuation Mark. CusurJump can take a binary value of "Y" or "N".

"p" is the position in the text box, "o" is the current text at that position, "n" is the change made to that position, and "t" is the time elapsed since the start of writing. In this example, the writer inserted a chatacter "I" at position 1 in the text box at a timestamp of 0.57 seconds, computed relative to when the writing started (i.e., at 0 elapsed seconds). The overall behaviral process of text production can then presented by a sequence of keystroke records. More importantly, qualitative labels may be attached to characterize a keystroke record in terms of the type (e.g., insertation, deletion) and location (e.g., inside of a word, end of a sentence) of an action, along with the content and associated time stamp. For the hypothetical example given in Table 1 (for illutration purpose only), the writer spends 5.13 seconds to write a full sentence: "It is a fun day." During the process, the writer changed the choice of a word from "good" to "fun" evidenced by a sequence of the "delete" actions. Cursor location is tracked so that if the cursor moves suddenly to a different location away from the current location, the jump behavior can be detected.

As Table 1 indicates, keystroke logs allow the visible aspects of the text-production process to be precisely reconstructed and retrospectively replayed. Figures 2 and 3 demonstrate one approach to visualizing the dynamics of the text-production process by plotting the time elapsed (horizontal axis) against text length and cursor position (verticle axis). When the writer is appending or deleting text at the end of the text, the dashed purple line (text length) and the solid green line (cursor position) would converge; when the writer is making changes elsewhere in the text, the green cursor-line would diverge from the purple length-line. The gaps between the two lines can indicate the degree of the "jump" action. The length-line can go up or down indicating adding or removing of content. The small-scale zig-zag pattern in both figures suggests that both writers conducted a fair amount of quick fixes or local edits mostly on the word level (e.g., typo correction, word-choice revision, removing/adding punctuation marks) at the end of the text as they write. The writer in Figure 3 showed evidence of global-level editing behavior towards the end of the writing session, when the writer moved the cursor to different parts of the text to make changes, as can be seen from the relatively large gaps between the purple and green lines. This type of jump-edit behavior is rather absent in Figure 2 for which the writer showed a much more linear writing process.

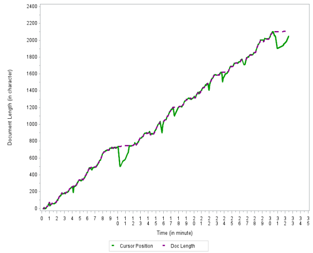

Figure 2: Writing Progression Example a

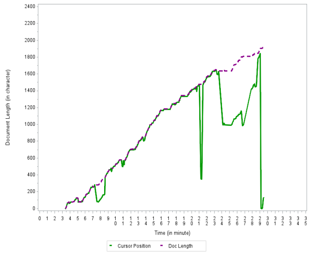

Figure 3: Writing Progression Example b

1.2 Inferences & Relations to Writing Quality

The nature and location of the changes that writers make to their text directly can directly support inferences about where the writer is cognitively in the composition process. For example, a long pause followed by insertion of a written outline is suggestive of task analysis and idea generation; a long pause at the phrasal or sentence boundaries followed by a burst of text production is a sign of sentence planning; alternating between insert and delete actions on a character-level inside of a word is likely an indication of spelling correction or word finding; if a writer types long sequences of words thereby adding new content, it is reasonably safe to assume that the writer is primarily engaged in content generation; and, if a writer jumps to various locations (tracked by cursor position) in the text to make changes, the writer is more likely in the state of text reviewing and revision (e.g., [6], [16]). Previous research has reported that the process of writing, such as the total time spent on writing, between-word pause tempo, initial pause length before typing a word, length of long burst (i.e., stretches of long sequence of text production), and extent of text editing and revision, relate to the quality of the final written product (e.g., [3], [17], [2], [27]). In this study, we largely followed the practice described in [8] by classifying each interkey interval (i.e., gap time betweeb two keystrokes) into one of the following four cognitive states in writing. These states intend to operationalize the theoretical subprocesses proposed in Hayes model, although there will unavoidably be gaps between theory and practice. *Long Pause* state, representing text planning, idea generation and deliberation, or hesitation and struggle with text production; *Text Production* state, representing relatively fluent content generation without major interruption where interruption is signaled by an extended pause; *Local Editing* state, representing localized (mostly on the word-level) minor text editing; and *Global Editing* state, representing reviewing, revision and copy editing on the whole passage/text level.

2. METHOD

2.1 Data Set

The data set was collected from a high-school equivalency testing program, which contains five subject tests: English language arts – reading, English language arts – writing, math, science, and social science. The focus of this study was the essay writing task in the writing subtest. In responding to the essay task, the examinees are expected to read two sources with different perspectives on a common issue (e.g., whether success is more the result of talent or of hard work), and then express and explain their opinions in writing while appropriately incorporating evidence from the sources. Each submitted essay was scored holistically on a 0-6 scale by two trained human raters according to a standardized grading rubric. Essays receiving a human score of 0 were excluded from analysis as those essays tend to have aberrant characteristics such as being empty, not in English, or consisting of random keystrokes. In this study, we selected two writing forms administered between September 2017 and August 2018 for investigation. Each form contained one essay writing task, or prompt. The sample size used for analysis was approximately 500 in each prompt. The analyses were conducted on the first (base) prompt, and then replicated on the second prompt to validate the consistency of the findings.

2.2 Propensity Score Matching

To ensure comparability, propensity score matching (PSM) [1] was used to minimize irrelevant factors such as performance level and the paraticipants' demographics and to balance covariates between the participants who responded to either of the two prompts. Also, the two different prompts were not administered at random, thus necessitating this step. A logistic regression model was developed to generate the propensity scores, and a one-to-one greedy matching without replacement algorithm with a caliper value of 0.05 was applied to find the matches in the prompt 2 sample to make it most comparable to the prompt 1 sample [15]. The caliper value refers to the maximum distance in propensity scores; hence the smaller the caliper value, the closer the match. In performing the PSM, the covariates were chosen based on our understanding of the examination and the examinee population, and on findings from previous reports on subgroup differences in writing process (e.g., [7], [25]). The covariates included for propensity score matching were gender (Male or Female), ethnicity (White, Black, Hispanic, or others/unreported), employment status (Full-time, Part-time, Unemployed, or others/unreported), highest education level (Below Grade 9, Some high school, others/unreported), English as best communicative language (Yes or No), as well as scores on the subject tests other than writing. All the demographic background variables were self-reported by the participants on a voluntary basis.

2.3 Feature Extraction

Keystroke logs were recorded automatically as writers composed their essays. A two-stage procedure was applied for feature extraction. In Stage 1, we classified each interkey interval (IKI) into one of four heuristically-defined and mutually-exclusive writing states by following the practice in [8] with some modifications. In Stage 2, we split each log into ten time periods by evenly dividing the total writing duration into ten segments, as one way to align and compare the logs of different length in duration. The choice of ten time-periods was made to balance the duration of each segment, which should be long enough to detect any patterns related to time distribution, *and* the number of segments, which should be sufficient for detecting longitudinal patterns.

Stage 1: Classification of writing states. The general programming logic is as follows.

- Step 1. Define Long Pause (LP) state. If an IKI is longer than n times in-word typing speed, it is then labelled as P. The keystroke sequence in between two adjacent Ps is considered a burst.
- Step2: Define Text Production (TP) state. Inside of a burst, if there is an absence of Delete action, or the max number of a Delete sequence is smaller than k, label all IKIs in this burst as TP. If there is a consecutive delete action sequence with k or more number of Deletes, temporarily change all IKIs in this burst to R, which will be refined in the next step.
- Step 3. Define Local Editing (LE) state. Use an m-IKI moving window to scan through the keystroke sequence within an R-burst. That is, the first moving window contains the 1^{st} to the m^{th} IKIs in a burst; the second moving window contains the 2^{nd} to the $(m+1)^{\text{th}}$ IKIs in the burst; and so on.
  - If all IKIs in a moving window are Inserts or contain less than s Deletes, change the first record in the moving window from R to TP. Continue with the same logic to the next moving window.
  - If a moving window contains equal to or more than s Delete actions, label all the m records in the moving window as LE.
- Step 4. Define Global Editing (GE) state. GE is indicated by text deletion while crossing sentence boundaries or making jump-edits elsewhere in the text away from the current location.

- If d or more consecutive Delete actions contain the following punctuation marks – comma, period, semicolon, exclamation mark, and question mark – label the entire Delete sequence as GE.
- If the position of an IKI in the text is different from the one before it by more than q, then the IKI is labeled as Jump Back (JB to an earlier place in the text) or Jump Forward (JF to a later place). Find the longest JB-JF pair in distance, and label all the IKIs in between to GE.

The parameters n, m, k, s, d, and q used in state definitions are customizable. In this study, we chose n=10, m=3, k=3, s=3, d=2, and q=4 as a starting point, mainly following [8]. Once all the IKIs are labelled, consecutive IKIs of the same state can be further aggregated. Each keystroke log can be described by the total number of states, the duration of each state, proportion of time spent on each state, the frequency of various transitions, etc. Among the four states, there are 12 state-transition possibilities (e.g., TP → GE) in total.

Stage 2: In this stage, we used the state classification generated above to calculate a writer's time distribution at various points during the writing process. To accomplish this goal, we divided each keystroke log into ten even time-periods. We then calculated the proportion of time spent in each writing state within a time period. As a result, each keystroke log was represented by a vector of 40 elements (i.e., four states times ten segments). The elements' values (i.e., percentages) could range from 0 to 100. When it is a 0, it simply means that the writer did not spend time on an activity (e.g., Local Editing) during that time period (which is one tenth of the total time). Similarly, when the value is 100, it means that a writer spent all his/her time on an activity (e.g., Global Editing) during that time period. With this information, the longitudinal pattern of time allocation can be revealed and investigated. We can analyze, for example, how writers spend their time at the beginning, middle, or end of their writing process, whether writers distribute their effort evenly or differently at various time-points during the writing process, and whether there are distinct profiles with regard to time management of an individual writing process.

2.4 Cluster Analysis and Interpretation

Using the writing samples in each of the two prompts, we conducted hierarchical cluster analysis (agglomerative approach) with Ward linkage using the Euclidean distance metric [23, 20]. The proportion of time spent in each state at the ten time-points was used to create 40 input variables. Each input variable was standardized to a mean of zero and standard deviation of one across individuals in a prompt [13]. The cluster analysis was done separately for each prompt, with the second prompt serving as a replication sample to verify results from the base prompt. Because there is no established convention for choosing the number of clusters, we used the Pseudo-F statistic, model R-squared, and semipartial R-square statistics to help us determine the appropriate number. The pseudo-F statistic is calculated as the ratio of the between-cluster variance to the within-cluster variance [14]. Larger values indicate better separation between the clusters. The model R-squared indicates the proportion of variance accounted for by the clusters. The semipartial R-square indicates the decrease in the proportion of variance accounted for due to joining two clusters. We plotted these statistics against the number of clusters to examine the impacts of joining or splitting clusters. Dendrograms visualizing the distances between the keystroke logs were also examined to help select the final number of clusters.

To interpret identified clusters, we compared how writers falling into different clusters allocated their efforts during the writing process. Since the proportions of time spent on Text Production, Local Editing, Global Editing, and Long Pause were used as input variables in the cluster analysis, this comparison would be the most direct way to examine any distinct patterns exhibited by each cluster. It would also be informative to know whether clusters are associated with distinct patterns of writing proficiency or in cluster members' demographic background. Therefore, to further substantiate the meaning of identified clusters, we compared the essay scores, essay length (in words), time on task (in minutes), the proportion of cluster members belonging to specific demographic categories, as well as a rough measure of writing efficiency calculated as essay length divided by time on task, between the clusters.

3. RESULTS

3.1 Outcome of Propensity Scoring Matching

The samples resulting from the propensity score matching are closely comparable between the two prompts (Table 2). In the first (base) prompt, males and females were evenly distributed; 53% of the examinees self-identified as White, 12% as Black, 14% as Hispanic; 4% reported that their highest grade level was below Grade 9, 62% reported having some high school education; 16% were working part-time, 17% were working full time, 23% were unemployed at the time of the examination. The majority of the examinees, 94%, indicated English as their best communicative language. The demographic background distribution of the second prompt, after matching, was very similar to that for the base prompt. The matched samples for prompt 2 also showed comparable means of the subtest scores to those for the base prompt.

3.2 Cluster Analysis Results

To decide the optimal number of clusters, we first examined the dendrograms, which indicated a solution of 3 or 4 clusters for both prompts. Figures 4 and 5 show the results of the Pseudo-F statistic and Sempartial R-square statistic that were considered for model selection in prompt 2. The results for prompt 1 are similar. The X-axis in both plots is the number of clusters ranging from 1 to 50. The Y axis is the Pseudo-F statistic on the left plot, R-squared on the middle plot, and semipartial R-square on the right plot. The results suggested a peak on the Pseudo-F statistic with a 3-cluster solution. The semipartial R-square plot shows an elbow point at cluster 3. The model R-squared (not shown) appears to go up continuously without a clear turning point. Considering all the evidence, we decided on three clusters as the most parsimonious and sensible solution for the this study sample.

3.3 Comparing the Clusters

Given the multivariate nature of the clustering variables, we drew radar charts to visualize the time distribution at the

Table 2: Comparability between Prompts after Propensity Score Matching

	Proportion										Mean Score (Scale: 0 - 20)			
Prompt	Female	White	Black	Hisp.	Below9	HS	PT	FT	Unemp.	Eng(Y)	Reading	Math	Science	SS
1 (base)	0.50	0.53	0.12	0.14	0.04	0.62	0.16	0.17	0.23	0.94	12.23	11.55	14.11	13.41
2 (matched)	0.51	0.52	0.13	0.14	0.02	0.63	0.16	0.19	0.20	0.94	12.19	11.73	13.97	13.55

Note: Below9: Below Grade 9; HS: some high school; PT: part-time employee; FT: full-time employee; Unemp.: unemployed; Eng(Y): English as best communicative language; SS: Social Science

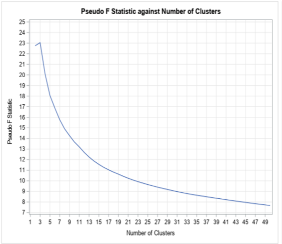

Figure 4: Pseudo-F statistic x N of Cluster

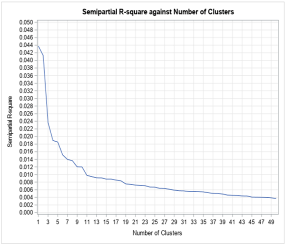

Figure 5: Semipartial R^2 x N of Cluster

ten consecutive time-points during the writing process of the logs belonging to each of the three clusters (Figures 6 and 7). The axes represent the average proportions of time spent on a state at a certain time. For example, "TP_1" refers to the proportion of time spent on Text Production at the beginning of writing – the 1^{st} of the ten duration segments. It is clear from Figures 6 and 7 that the three clusters demonstrated rather different polygonal shapes over all axes, which are consistent between the two prompts.

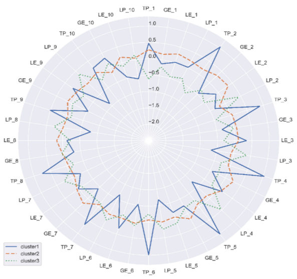

Figure 6: Radar Chart (Prompt 1)

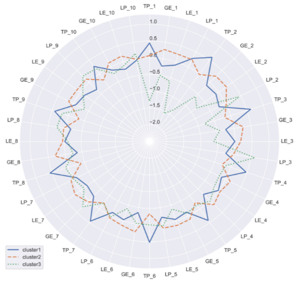

Figure 7: Radar Chart (Prompt 2)

Cluster 1 (blue colored, solid line in both prompts) writers have a distinct pattern with notable spikes on the TP state over the course the writing process. Because the total writing time is constrained and writers can only do one thing at a time, if the writers spend more time on text production, they would necessarily spend less time on the other activities such as editing or revision. This constraint is evident from the plots where, for Cluster 1, the proportion of time spent on long pauses and editing is relatively smaller. Cluster 2 (orange colored, dashed line) writers, on average, have a much smoother circle compared to Cluster-1. Cluster 2 writers appeared to have distributed their efforts more evenly throughout the writing process. The allocation of time across the four writing states is relatively balanced over the course of the writing session. In general, Cluster 1 seems to represent a group of writers that compose linearly without showing much editing behaviors, while Cluster 2 seems to represent writers that also consistently produce text but still spend time on text planning and conduct text editing and revision as they write. Cluster 3 (green colored, dotted line) writers further showed a distinct time-management pattern from the other two clusters. The writers in Cluster-3 appeared to have difficulties in generating text at the start of the writing session, as evidenced by the lack of text production (TP_1) and a higher proportion of the local editing behaviors (GE_1 and LE_1) in the first time period, which suggested possible false starts during the writing process. This "struggling" pattern appeared to have persisted into later stages of the writing process, as evidenced by a higher proprotion of time spent on long pauses compared to text production or editing.

We also examined the actual length of time writers stayed in a state before transitioning to a different state. The results suggest that the three clusters not only differ in their relative time distributions during writing, but also in the total time writers stayed in different states. Each cluster displayed distinct patterns consistent across prompts: Cluster 1 writers spent considerably less time on long pauses than Clusters 2 and 3 writers; Cluster 2 writers spent notably larger amounts of time making word-level local edits than Clusters 1 and 3 writers by approximately 1 minute in Prompt 1 and 2 minutes in Prompt 2; and Clusters 3 writers generally spent less time on text production than Clusters 1 and 2 writers by about 1-2 minutes in Prompt 1 and about 3.5 minutes in Prompt 2. An additional interesting difference between Clusters 1 and 2 is that Cluster 2 writers not only spent longer time on local editing, but also on global editing by about 3 minutes in both prompts. A close comparison between Clusters 2 and 3 further revealed that Cluster 2 writers also appeared to spend more time on long pauses than Cluster 3 writers by a small margin in Prompt 1 and by a rather large margin in Prompt 2.

Taking into account both absolute time spent in each state, and relative time in each state, Cluster 2 writers appear to have shown a stable-tempo, iteractive process pattern in which they switch repeatedly between the activities of text planning, text production, and text editing over the course of their writing sessions. Although more data are needed to verify this interpretation, the long pauses demonstrated by Cluster 3 writers throughout the writing session appeared to be signals of hesitation and difficulties in content generation. Finally, Cluster 1 writers seem to be relatively quick and fluent at generating ideas (as evidenced by fewer long pauses) and at translation and transcription (that is, at expressing their ideas in written form).

To better understand and interpret the identified clusters, Table 3 further compares the characteristics of student essays across the three clusters. Cluster 1 writers spent the shortest time on the writing task on average (22.94 minutes

Table 3: Proficiency and Demographic Distributions of Clusters

		Prompt 1			Prompt 2		
	Variable	Cluster 1	Cluster 2	Cluster 3	Cluster 1	Cluster 2	Cluster 3
	n	65	264	168	169	259	71
	Essay Score (scale: 2 - 12)	5.51 (1.70)	5.64 (1.54)	5.43 (1.53)	5.38 (1.59)	5.42 (1.74)	4.51 (1.80)
	Essay Length (in words)	305 (136)	324 (145)	289 (128)	287 (129)	290 (127)	236 (118)
	Time on Task (in minutes)	22.94 (11.56)	36.60 (15.65)	32.36 (14.16)	25.45 (12.39)	36.04 (14.86)	26.88 (14.37)
	Efficiency (words/min)	15.02 (6.19)	9.94 (4.43)	9.96 (4.09)	12.85 (5.71)	8.97 (3.98)	10.82 (6.41)
Proportion	Female	0.52	0.56	0.43	0.49	0.54	0.39
	White	0.62	0.51	0.50	0.58	0.51	0.46
	Black	0.12	0.14	0.13	0.12	0.12	0.13
	Hispanic	0.06	0.13	0.18	0.09	0.17	0.13
	Below Grade 9	0.03	0.02	0.02	0.05	0.03	0.03
	Some high school	0.54	0.63	0.64	0.60	0.62	0.68
	Part-time	0.15	0.13	0.20	0.17	0.16	0.14
	Full-time	0.12	0.19	0.22	0.17	0.17	0.20
	Unemployed	0.20	0.21	0.18	0.21	0.24	0.28
	English as Best (Y)	0.96	0.94	0.96	0.96	0.92	0.94

Note: Values in parenthesis are standard deviations. Values on the lower-half of the table are the percent of various subgroups within a cluster.

in Prompt 1; 25.45 minutes in Prompt 2). The difference in the total time on task between Clusters 1 and 2 is drastic – about 14 minutes difference in Prompt 1 and about 10 minutes difference in Prompt 2. Cluster 1 writers wrote notably more words/minute than Cluster 2 writers (15.02 vs 9.94 in Prompt 1; 12.85 vs. 8.97 in Prompt 2). The overall evidence seems to suggest Cluster 1 writers were more efficient than Cluster 2 writers, in that they spent significantly less time writing, yet achieved comparable text quality (essay scores).

In relation to demographic background, several results are noteworthy. On both prompts, Cluster 1 contained a notably greater proportion of White writers, a lower proportion of Hispanic writers, and a lower proportion of examinees with high-school experience, compared to the overall demographic distribution in Table 2. Cluster 2 included a slightly greater proportion of female writers than the average on both prompts, while all other demographic variables fell close to the mean for each prompt. Finally, Cluster 3 had a considerably lower proportion of White or female writers. But the results in general are less consistent between the two prompts for Cluster 3. The evidence seems to suggest that writers from different demographic background and having different educational experience may display distrinctive patterns in their writing processes. However, without further evidence, we cannot infer any causal connection between demographic group membership, writing process patterns, and overall writing performance.

4. DISCUSSION

In this paper, we presented a study on the use of keystroke analytics to understand writers' cognitive processes during writing. One possible outcome of this study, and of the larger research program of which it is a part, would be to providing actionable writing feedback to instructors and learners. However, before we can reach this goal, we need to develop a clear understanding of how writing processes change as a result of learning and instruction. This study provides a first attempt to address this issue, by identifying characteristic longitudinal patterns of time management that writers display when they respond to an essay writing task. The current results suggest that there are at least three distinct writing profiles that describe how writers approach an on-demand essay writing task. Though, it will be critical that the analysis to be replicated with more data. In the future, for writers placed into different profiles, we can imagine giving customized suggestions on writing strategies to improve learning and practice. Obviously, more research is needed to further validate the meaning and interpretation of the profiles we have detected. Possible approaches include cognitive interviews to elicit writers' understanding of what they were doing, combining eye-tracking technique with keystroke logging to get a better sense of where the writer's attention was focused and how changes in the focus of attention interacts with pause patterns, and convening an expert panel to determine whether the clusters derived by atatistical analysis align with expert judgments about what the writers were doing at each point in the writing process. The availability of keystroke logs makes it possible to replay a writer's composition process like a movie. Such replays can then be presented as stimuli to assist and guide cognitive interviews or the expert review process. Statistical tests will be necessary to detect if the profiles are significantly different beyond the practical importance. It will also be essential to replicate our analyses on a wide range of writing prompts, a broader variety of writing tasks, and across many different writer populations, as well as to study how well findings resulting from timed-writing tasks can be generalized to writing tasks with no time restriction.

The association between cluster assignment and demographic background is worth further investigation. The study of writing process ought to integrate with the social and linguistic context [21]. Previous studies have reported subgroup differences in writing processes (e.g., between native and non-native speakers in [24], between male and female writers in [26], between black and white students in [7]). It is concievably helpful and valuable to give information about writing profiles to provide customized feedback to writers from different linguistic, social, and educational backgrounds.

Finally, although this is beyond the scope of the current study, it is worth mentioning that keystroke-enabled process visualization such as those illustrated in Figures 2 and 3, in and of itself, may have instructional value, by making it easier for students to understand and self-reflect their writing processes. For instance, teachers may select replays and graphs to demonstrate a specific writing subprocess or a writing strategy, to help the class understand how to implement a more effective writing process. Teachers might also be able to use such graphs during their one-on-one conference with their students to help them better understand their writing strengths and weaknesses (such as lack of editing, low-level engagement, or lack of sufficient attention to idea generation) that are revealed by the keystroke log. Future research is encouraged to gather teacher and students feedback on the assistive value of keystroke logs in their teaching and learning experience.

5. REFERENCES

[1] P. C. Austin. An introduction of propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, 46(3):399–424, 2011.

[2] V. M. Baaijen, D. Galbraith, and K. de Glopper. Keystroke analysis: Reflections on procedures and measures. *Written Communication*, 29(3):246–277, 2012.

[3] I. Breetvelt, H. van den Bergh, and G. Rijlaarsdam. Relations between writing processes and text quality: When and how? *Cognition and Instruction*, 12(2):103–123, 1994.

[4] E. F. Burke. An experimental study of the educational use of the typewriter in second grade. *Master's Thesis*, Loyola University, 1939.

[5] P. Deane and M. Zhang. Automated writing process analysis. In D. Yan, A. Rupp, and P. Foltz, editors, *Handbook of automated scoring: Theory into practice*, pages 347–364. Chapman and Hall, 2020.

[6] D. Galbraith and V. M. Baaijen. Aligning keystrokes with cognitive processes in writing. In E. Lindgren and K. P. H. Sullivan, editors, *Observing writing*, pages 306–325. Brill Publishing, 2019.

[7] H. Guo, M. Zhang, P. Deane, and R. Bennett. Writing processes differences in subgroups reflected in keystroke logs. *Journal of Educational and Behavioral Statistics*, 44(5):571–695, 2019.

[8] H. Guo, M. Zhang, P. Deane, and R. Bennett. Effects of scenario-based assessment on students' writing processes. *Journal of Educational Data Mining*, 12(1):19–45, 2020.

[9] J. R. Hayes. Modeling and remodeling of writing. *Written Communication*, 29:369–388, 2012.

[10] M. Leijten and L. V. Waes. Keystroke logging in writing research: Using inputlog to analyze and visualize writing processes. *Written Communication*, 30:358–392, 2013.

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

[12] D. McPherson. A study of typing speed and accuracy development using computer-based and typewriter-based instruction in a public high school. *OTS Master's Level Project & Papers*, 352, 1995.

[13] G. W. Millgan and M. C. Cooper. A study of standardization of variables in cluster analysis. *Journal of Classification*, 5:181–204, 1988.

[14] G. W. Milligan and M. C. Cooper. An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 502:159–179, 1985.

[15] T. Nguyen, G. S. Collins, J. Spence, J. Daures, D. P. J., P. Landais, and Y. L. Manach. Doubling-adjustment in propensity score matching analysis: Choosing a threshold for considering residual imbalance. *BMC Medical Research Methodology*, 17, 2017.

[16] T. Quinlan, M. Loncke, L. M, and L. V. Waes. Coordinating the cognitive processes of writing: The role of the monitor. *Written Communication*, 29(3):345–368, 2012.

[17] S. Sinharay, M. Zhang, and P. Deane. Prediction of essay scores from writing process and product features using data mining methods. *Applied Measurement in Education*, 32(2):116–137, 2019.

[18] C. K. Stallard. An analysis of the writing behavior of good student writers. *Research in the Teaching of English*, 8:206–218, 1974.

[19] K. P. Sullivan and E. Lindgren. *Computer keystroek logging and writing: Methods and applications.* Elsevier, New York, 2006.

[20] G. J. Szekely and M. L. Rizzo. Hierarchical clustering via joint between-within distances: Extending ward's minimum variance method. *Journal of Classification*, 22(2):151–183, 2005.

[21] L. V. Waes and P. J. Schellens. Writing profiles: The effect of the writing mode on pausing and revision patterns of experienced writers. *Journal of Pragmatics*, 35:829–853, 2003.

[22] S. Wallot and J. Grabowski. Typewriting dynamics: What dintinguishes simple from complex writing tasks. *Ecological Psychology*, 25:1–14, 2013.

[23] J. H. Ward. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244, 1963.

[24] C. Xu and Y. Ding. An exploratory study of pauses in computer-assisted efl writing. *Language Learning & Technology*, 18:80–96, 2014.

[25] M. Zhang, R. Bennett, P. Deane, and P. van Rijn. Are there gender difference in how students write their essays? an analysis of writing processes. *Educational Measurement: Issues and Practice*, 39(2):14–26, 2019.

[26] M. Zhang, R. E. Bennett, P. Deane, and P. van Rijn. Are there gender differences in how students writer their essays? an analysis of writing processes. *Educational Measurement: Issues and Practice*, 38(2):14–26, 2019.

[27] M. Zhang, J. Hao, C. Li, and P. Deane. Classification of writing patterns using keystroke logs. In L. A. van der Ark, D. M. Bolt, W.-C. Wang, J. A. Douglas, and M. Wiberg, editors, *Quantitative psychology research: The 80th Annual Meeting of the Psychometric Society, Beijing, 2015*, pages 299–314. Springer, 2016.