Article

Journal of Educational and Behavioral Statistics
2019, Vol. 44, No. 5, pp. 571–596
DOI: 10.3102/1076998619856590
Article reuse guidelines: sagepub.com/journals-permissions

Writing Process Differences in Subgroups Reflected in Keystroke Logs

Hongwen Guo
Mo Zhang
Paul Deane
Randy E. Bennett
Educational Testing Service

We used an unobtrusive approach, keystroke logging, to examine students' cognitive states during essay writing. Based on data contained in the logs, we classified writing process data into three states: text production, long pause, and editing. We used semi-Markov processes to model the sequences of writing states and compared the state transition time and probability for demographic subgroups that were matched on writing proficiency. Results suggested that the subgroups employed different processes in essay writing.

Keywords: *process data; writing proficiency; semi-Markov model*

1. Introduction

A major attraction of keystroke logging is that the moment-by-moment process of text creation is recorded unobtrusively, allowing the text production process to be analyzed more naturally. From these logs, we can extract observables indicating various states of the writing process, where long pauses may suggest a planning state, text deletion and out-of-order insertion may suggest an editing state, and rapid uninterrupted text bursts may denote a text production state. These logs may help researchers understand cognitive writing processes in general as well as how individuals and demographic groups differ in their approaches to composition. The information provided by such logs goes well beyond what can be discerned from the final essay. Ultimately, researchers hope that the keystroke logs may provide informative advice to individual writers to improve writing quality.

Previous research on keystroke logs has found a variety of interesting results. For example, experienced and unskilled writers differ in their planning and revision patterns. More skilled writers mainly do conceptual revision (e.g., revise words related to the meaning and content of the text), while unskilled writers show more local corrections to punctuation, syntax, and spelling (Breetvelt, van den Bergh, & Rijlaarsdam, 1994; McCutchen, 1996; Hayes, 2012). In addition, compared to novice writers, more experienced writers tend to pause longer at natural text junctures (Matsuhashi, 1981; Schilperoord, 2002), reflecting engagement in higher level processes such as idea generation (Schilperoord, 2002). Zhang, Bennett, Deane, and van Riin (2019) found gender differences in essay writing as well.

Besides describing planning and revision patterns of skilled versus unskilled writers, keystroke logging research has examined the meaning and stability of different log features. Deane and Zhang (2015) found that the text burst length, an indicator of fluency, had a strong association with essay quality. Allen et al. (2016) found that keystroke indices accounted for 38% of the variance in the linguistic characters. Finally, Guo, Deane, van Rijn, Zhang, and Bennett (2018) discovered that the interkey interval in conjunction with the total writing time explained as much as 47% of the variance in writing scores; in addition, some keystroke features were quite consistent across essays written by the same students in response to parallel prompts.

A potentially valuable way to view writing processes is in terms of states. Kellogg (2001) classified these states into planning, translation, and revision, which compete for a limited common, general-purpose cognitive resource. Planning content involves idea generation and text organization. Translation includes the linguistic and motoric operations needed to generate sentences from the planned content. Revision involves reviewing and amending the text while detecting and correcting errors or problems either in the text or in the plan for the text. Subsequent to Kellogg, Hayes (2012) proposed a somewhat different conceptualization in which he distinguished planning, translating, and reviewing, but as subcomponents of a "transcription" state or process.

Other researchers have used keystroke logs to infer these states and to attempt to decompose them. For example, Baaijen, Galbraith, and Glopper (2012) focused exclusively on the procedures and measures to analyze pauses, bursts, and revisions, the basic units of analysis from keystroke logs in terms of cognitive models of writing. A principal components analysis identified three underlying dimensions in these data: planned text production, within-sentence revision, and revision of global text structure. A tagging process using think-aloud protocols (Kaufer, Hayes, & Flower, 1986, pp. 125–126) was also employed. More recently, Zhang and Deane (2015) used principal factor analysis and found a four-factor structure of 29 writing-process features extracted from keystroke logs. That structure included general fluency (an aspect of translation), major editing, local editing, and planning and deliberation.

The goals of the current study are to use keystroke logs to classify students' writing processes into sequences of writing states and then use semi-Markov processes to model these sequences. In addition, we investigate subgroup differences in those writing processes, independent of writing proficiency. To address these research questions, we conducted three analytical steps. Step 1 is classification. Based on the previous studies, sequences of keystroke actions are classified into three states: P (long pause), E (editing, a subcomponent of revision), and T (text production). Long pauses are most saliently connected to planning or reviewing the text produced so far; they might also reflect other situations such as disengagement or students being frustrated or stymied because they are unable to generate content easily. Editing mostly includes out-of-order insertion and deletion. Finally, text production entails mostly uninterrupted text-generating bursts. This classification is similar to an automated annotation process in that it does not require manual tagging by writers, investigators, or other instruments as is the case for think-aloud or eye-tracking protocols (Leijten & Van Waes, 2013).

Step 2 is score matching. Using a statistical matching method, studied demographic subgroups are matched on essays scores to eliminate possible confounding of group writing proficiency-level differences with differences in group writing processes. By controlling essay scores, the differences in writing processes are more likely to be explained by the group characteristics instead of by writing proficiency.

Step 3 is process modeling. Because we are interested in the writing states (a categorical variable) and the time duration at each state (a continuous variable), we use the semi-Markov model (Krol & Saint-Pierre, 2015) that generalizes the multistage and continuous-time Markov chain model (CTMC; Jackson, 2011). Comparisons are made between subgroups in terms of how likely a typical writer is to move from one state to another and how long on average that writer stays at the current state (the sojourn time) before transiting to the next one. In addition, we compute summaries of keystroke features at the essay level to assist with interpretation of the semi-Markov results.

2. Method

2.1. Data

The data set came from a larger experimental study (Zhang, van Rijn, Deane, & Bennett, 2019) in which a base summative writing assessment form was administered along with three alternative forms varying in task order and/or topical focus. The sample from the base test form was used for analyses in this study. In the assessment, students were asked to complete a sequence of items related to the topic of whether advertisements directed at children should be banned and then write an argument essay on that topic. In the final data set, there were 257 eighth graders who had valid keystroke logs on the essay task, along with human scores for essay quality. Students were asked to finish their essays within 30 minutes.

The essays were scored against two rubrics, each on an integer scale ranging from 0 to 5. For the purpose of our analyses, we excluded responses that received a human score of 0 (a very small subset of the total), since that score point was used to denote essays with unusual response characteristics including empty, nonsensical, and off-topic responses; plagiarized responses; and responses consisting of random keystrokes. One of the scoring rubrics (denoted as RS1) evaluated basic writing skills (e.g., word usage, writing mechanics, syntactic variety, grammar, text organization), and the other rubric (denoted as RS2) evaluated student performance on such higher level skills as the quality of the argument. Human scores were computed for each rubric as a mean score taken across two raters, except when raters disagreed by more than one point, in which case a third rater was employed to resolve the discrepancy. We used the human scores as criterion variables in this study.

2.2. Writing States

To use a semi-Markov model, it is necessary to classify the keystroke logs into different states. For that, we use the following features in the raw keystroke data (Deane & Zhang, 2015):

- The raw keystroke logs are first grouped into chunks. Most of the time, the unit of a *chunk* is a word or a delimiter. A chunk may also be comprised of such text segments as sequences of delimiters, deleted word sequences, or inserted word sequences, among others.
- A *burst* is defined as a sequence of chunks without interruptions by long pauses. A burst break (or a *long pause*) is calculated continuously on the fly as writing proceeds for each writer in a way that it is adaptive using 4 times the median between-chunk pause length of all chunks thus far. A burst is typically comprised of more than one word, also called a phrasal burst.
- *Between burst* is a chunk of an isolated event (most likely a space, line or paragraph break, or comma or period at phrasal or sentence boundaries), accompanied by long pauses before and after it. A between burst is considered as indicative of *long pause* in this study. We further treat the initial pause at the beginning of a burst (including a one-word burst) as a meaningful pause indicating a *long pause* state. A special case arises when a chunk occurs between bursts and is comprised of only an alphanumeric word. In this case, the chunk is labeled as a *one-word burst*—a special burst case.
- Finally, if a burst contains deleted sequences, replaced sequences, out-of-order inserted sequences, or other major editing (e.g., typo corrections of more than two characters), it is classified as an *editing* state; otherwise, it is denoted as a *text production* state.

Duration/holding time for each state is captured using the interkey intervals in the keystroke logs in the time unit of 1 second. Generally speaking, the three states (long pause, text production, and editing) correspond to proposer/planner, translator and transcriber, and evaluator in Hayes's (2012) cognitive writing process model.

Because of the nature of chunk data, the same states may appear consecutively. In the analysis, we combine the adjacent, same states into one. As a result, the

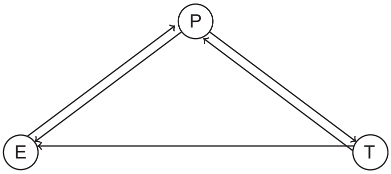

FIGURE 1. *Transition relations between writing states. P = long pause; E = editing; T = text production.*
Source. Copyright by Educational Testing Service, 2019. All rights reserved.

transition probability of a state to itself is always zero. This decision leads to the triangular structure of the state relationships in Figure 1. Note that State E cannot immediately move to State T because there is always an initial pause P before T.

We applied the above classification approach to each student's keystroke log sequence. In our data set, there are 374 chunks on average across all 257 valid keystroke log sequences produced by these students ($N = 257$), with the largest chunk number being 918 and the smallest having 1 chunk only. We removed logs that were too short ($\leq$ 50 chunks) and retained 231 log sequences (i.e., we retained 231 students' keystroke logs).

Figure 2 shows the state sequences of three students. The x-axis is time in seconds, and the y-axis is writing state. In the upper panel, the student's total number of states is 113, and his or her RS1 score is 1; in the middle panel, the student's total number of states is 161, and his or her RS1 score is 3; and in the lower panel, the student's total number of states is 136, and his or her RS1 score is 4.

2.3. Subgroups

Students' background variables—gender, race, socioeconomic status (SES; whether students received free or reduced-price lunch)—and their writing scores assigned by human raters on both rubrics were obtained. Table 1 shows the mean writing scores for each subgroup. As expected, low SES students had lower essay scores compared to their peers, the relatively high SES students; males had lower essay scores compared to females; and Black students[1] had lower scores than White students. Analysis of variance (not shown) revealed that gender, race, and SES were associated with writing scores, a result also found in the National Asessment of Educational Progress (National Center for Education Statistics, 2011).

2.4. Semi-Markov Model

Markov chains or Markov models are often used in educational data analyses and other fields. Nevertheless, they may not always fit the data well, and more generalized models may be preferable. In the current study, we applied the semi-Markov models to our data.

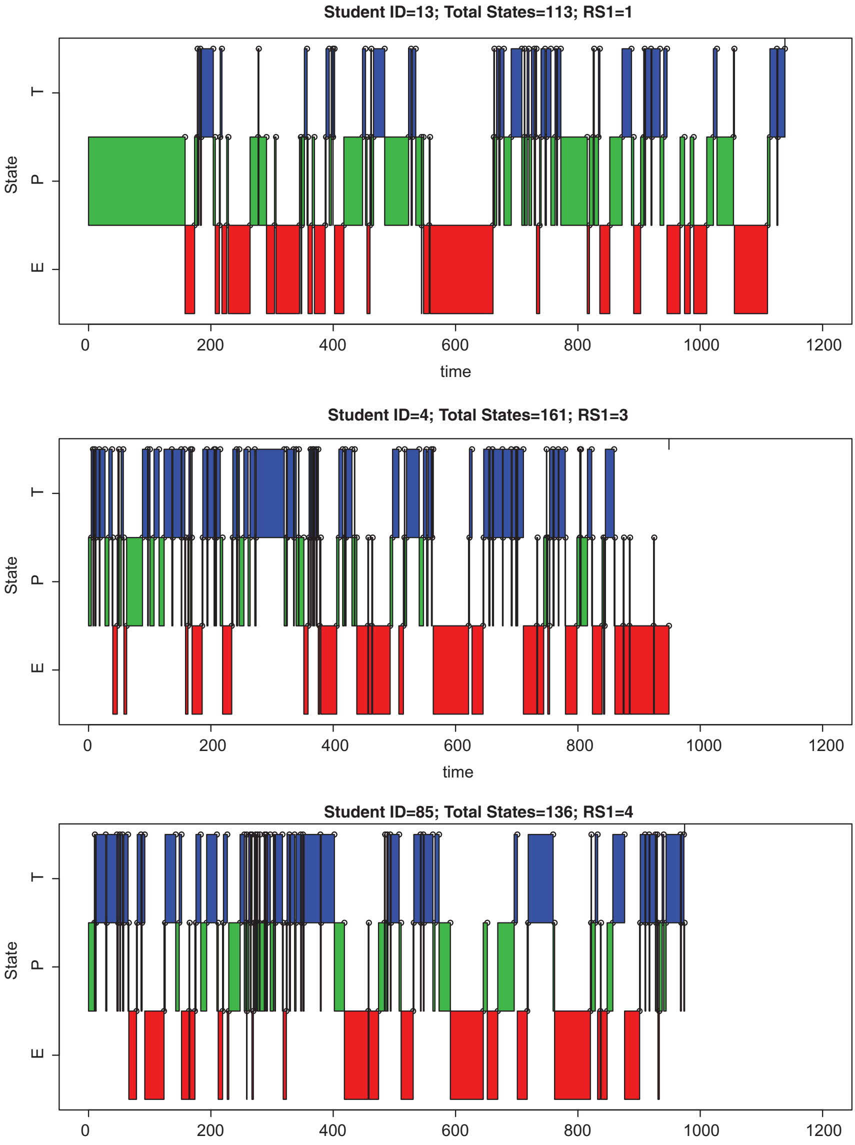

FIGURE 2. *State sequences of three students. The* x-*axis stands for time in seconds, and the* y-*axis for state of E (editing), P (long pause), and T (text production).*
Source. Copyright by Educational Testing Service, 2019. All rights reserved.

TABLE 1.
Summary Statistics for Human Scores for Each Subgroup

	SES			Gender			Race		
Score	Group	*n*	Mean (*SD*)	Group	*n*	Mean (*SD*)	Group	*n*	Mean (*SD*)
RS1	Low	52	1.69 (1.00)	Female	110	2.30 (1.13)	Black	23	1.35 (0.65)
	High	179	2.27 (1.01)	Male	121	1.99 (0.93)	White	208	2.23 (1.04)
RS2	Low	52	2.04 (1.12)	Female	110	2.65 (1.17)	Black	23	1.65 (0.98)
	High	179	2.60 (1.14)	Male	121	2.31 (1.13)	White	208	2.56 (1.14)

Note. *SD* = standard deviation. RS1 = scoring rubrics 1; RS2 = scoring rubrics 2.
Source. Copyright by Educational Testing Service, 2019. All rights reserved.

The states in a Markov chain are characterized either by discrete time points or by a continuous time variable. The discrete-time Markov chain and CTMC (Jackson, 2011) both require the memoryless property that the probability of a future state depends only on the current state, and, consequently, the interarrival time (also called sojourn time, duration, or holding time) is either fixed or follows an exponential distribution. In cases when this assumption is too restrictive, semi-Markov models can be considered (Krol & Saint-Pierre, 2015).

To characterize a semi-Markov model, two so-called hazard rates are necessary. One is the hazard rate of duration time (or sojourn time, denoted as $\alpha_{ij}(t)$, which is the likelihood of leaving the current state *i* for the next state *j* at time *t*); a larger α indicates a shorter duration time. The other is the hazard rate of the process (or the transition intensity, instantaneous transition probability, denoted as $\lambda_{ij}(t)$, which is the probability of transition from state *i* to state *j* at time *t*); a larger λ indicates a more frequent transition. Details can be found in the Appendix.

When the distribution of the sojourn time follows the exponential distribution that has one parameter $1/\sigma_{ij}$, the semi-Markov process is a CTMC, and the hazard rate of the sojourn time $\alpha_{ij}(t) = 1/\sigma_{ij}$ is constant; a larger hazard rate of α_{ij} implies a shorter average stay at the current state (*i*) before entering next state (*j*). When the distribution of the sojourn time follows the Weibull distribution that has a shape parameter ν and a scale parameter σ, the hazard rate $\alpha_{ij}(t) = \nu_{ij}/\sigma_{ij} \times (t/\sigma_{ij})^{\nu_{ij}-1}$ is a function of time *t* (when $\nu = 1$, the Weibull distribution degenerates to the exponential distribution; when $\nu < 1$, the hazard rate function decreases in the sojourn time; and when $\nu > 1$, the hazard rate function increases with the sojourn time). The hazard rate of $\lambda_{ij}(t)$ is more complicated for the semi-Markov models and is discussed in the Results section.

2.4.1. Cox regression model. Our study focuses on subgroup comparison. Therefore, a covariate *Z* associated with the studied groups was used in the

Cox proportional regression model (Cox, 1972) to compare group processes. The influence of Z is placed on the hazard rate $\alpha_{ij}(t)$ in the semi-Markov model by

$$\alpha_{ij}(t|Z) = \alpha_{ij0}(t)\exp(\beta_{ij}Z), \tag{1}$$

where α_1 and α_0 are the hazard rates of the focal ($z = 1$) and reference ($z = 0$) groups, respectively. If the proportional hazard assumption holds, a hazard ratio of one (i.e., $\beta = 0$) means equivalence in the hazard rates of the two groups (i.e., $z = 1$ vs. $z = 0$), whereas a hazard ratio other than one indicates a difference between groups. A likelihood ratio test is used to determine whether the regression coefficient is statistically different from zero. Under the null hypothesis, the statistic has an approximate χ^2 distribution with one degree of freedom.

In this study, we use the R packages *SemiMarkov* (Krol & Saint-Pierre, 2015) for semi-Markov models to estimate related parameters, hazard rates $\lambda_{ij}(t)$, $\alpha_{ij}(t)$, and the covariate effect β_{ij}. The maximum likelihood estimation method can be used to obtain estimators in the semi-Markov process.

2.5. Matching Method

Because we are interested in evaluating subgroup differences in writing processes that may be attributed to group writing styles, instead of to differences in writing proficiency, we match the focal and reference subgroups on their writing scores. Using the original writing scores as a covariate in Equation 1 is possible only if we dichotomize the scores, which would lead to coarser matching. Exact matching would eliminate more process data given our limited data pool. Hence, we decided to employ the commonly used matching procedure, the propensity score method (Rosenbaum & Rubin, 1983), to preserve as much data as possible. Using the R package *MatchIt* (Ho, Imai, King, & Stuart, 2007), we estimate the propensity score of each student in both studied subgroups, with RS1 as a covariate, and then students in the reference group are selected and matched to those in the focal group, based on the method of choice. We used the “nearest” method in our implementation. This matching procedure is applied to prepare the data before fitting the semi-Markov model for subgroup comparison.

3. Results

In this study, we used the R language (R Core, 2018) to conduct statistical analyses and to produce graphs.

3.1. Total Group

In this section, we present results of model fit analysis and hazard rate estimation for the total group. Table 2 displays the empirical transition frequencies of the writing states across all students, where E, P, and T stand for editing, long

TABLE 2.
Empirical State Transition Table

	Editing	Long Pause	Text Production
E	*4,665*	4,573	—
P	3,809	*9,267*	5,381
T	831	4,488	*5,381*

Source. Copyright by Educational Testing Service, 2019. All rights reserved.

TABLE 3.
Estimated Parameters of Weibull Distributions

Label	Transition	Estimate	*SD*	LCI	UCI	H_0	Statistic	*p* Value
σ	E → P	21.93	.40	21.13	22.72	1.00	2,688.11	<.0001
σ	P → E	4.50	.11	4.29	4.71	1.00	1,065.31	<.0001
σ	P → T	4.23	.08	4.08	4.39	1.00	1,629.32	<.0001
σ	T → E	6.41	.26	5.90	6.92	1.00	429.99	<.0001
σ	T → P	2.53	.03	2.46	2.60	1.00	1,987.51	<.0001
ν	E → P	0.99	.01	0.97	1.01	1.00	0.33	.5657
ν	P → E	0.72	.01	0.71	0.74	1.00	1,546.41	<.0001
ν	P → T	0.77	.01	0.75	0.78	1.00	1,093.47	<.0001
ν	T → E	0.91	.02	0.86	0.95	1.00	17.46	<.0001
ν	T → P	0.69	.01	0.67	0.70	1.00	1,640.85	<.0001

Note. σ = scale parameter; ν = shape parameter. LCI = lower confidence interval; UCL = upper confidence interval; *SD* = standard deviation.
Source. Copyright by Educational Testing Service, 2019. All rights reserved.

pause, and text production. The numbers on the diagonal are the total numbers of E, P, and T in our data set, which are 4,665, 9,267, and 5,381. The off-diagonal numbers are the counts of transitions. For example, 3,809 (cell in the second row and the first column) is the count of transitions from P to E across all students. As noted before, in our sample, there were 231 students in total.

Table 3 shows the estimated parameters of the semi-Markov model using all data, where *SD* is the estimation error, LCI and UCI stand for the lower and upper points of the 95% confidence interval, and the last three columns are related to the Wald test: the null hypothesis of H_0, test statistic, and *p* value.

In Table 3, all σ estimations are significantly larger than 1, indicating that the sojourn time has larger spread than 1. Of these estimations, the sojourn time from State E to P is the most dispersed. The estimated νs are not significantly different from 1 for the E-to-P transition, indicating that the sojourn time may be approximated by an exponential distribution; that is, the

TABLE 4.
Estimated Parameters of Weibull Distributions

Label	Transition	Estimation	*SD*	LCI	UCI	H_0	Statistic	*p* Value
σ	E → P	16.03	.24	15.56	16.49	1.00	4,020.25	<.0001
σ	P → E	4.50	.11	4.29	4.71	1.00	1,063.13	<.0001
σ	P → T	4.24	.08	4.08	4.40	1.00	1,635.25	<.0001
σ	T → E	6.24	.25	5.75	6.74	1.00	427.14	<.0001
σ	T → P	6.13	.10	5.94	6.33	1.00	2,697.62	<.0001
ν	P → E	0.72	.01	0.71	0.74	1.00	1,540.33	<.0001
ν	P → T	0.77	.01	0.75	0.78	1.00	1,085.85	<.0001
ν	T → E	0.91	.02	0.86	0.95	1.00	16.94	<.0001
ν	T → P	0.98	.01	0.96	1.00	1.00	3.71	.0541

Note. σ = scale parameter; ν = shape parameter in the adjusted semi-Markov model (i.e., $\nu_{12} \equiv 1$). LCI = lower confidence interval; UCL = upper confidence interval; *SD* = standard deviation.
Source. Copyright by Educational Testing Service, 2019. All rights reserved.

hazard rate of the sojourn time may be approximated by a constant for this transition; in this case, the duration time at E before P is memoryless. All the other estimated νs are significantly less than 1, indicating that the sojourn time may not be approximated by exponential distributions, and the hazard rate of the sojourn time decreases with time, a diminishing likelihood of ending the current state as the sojourn time lasts longer. Therefore, Weibull distributions were assumed for these transitions.

Table 4 shows the estimated parameters for the semi-Markov model with $\nu_{12} \equiv 1$. The overall patterns in Table 4 are similar to those in Table 3. Figure 3 shows the estimated density distributions of sojourn/duration time for the transitions in Table 4 with fixed ν_{12} in the model compared to empirical data; density distributions with free ν_{12} are somewhat similar and are not presented. The figure shows that the fit of estimated distributions is reasonable, except for E-to-P, which may call for distributions other than exponential or Weibull distributions. The sojourn times have a very skewed distribution with a long tail on the right; the median duration at E before entering P (i.e., transition E-to-P) is 15.16 seconds; the median duration for the other transitions is short and within 5 seconds.

Figure 4 shows the estimated hazard rates of sojourn time for these transitions with the constraint $\nu_{12} \equiv 1$. The y-axis is the hazard rate, and the x-axis is sojourn time in seconds. Figure 4 shows that the hazard rates decrease with the sojourn time with the exception of E-to-P, which is constant because an exponential distribution was assumed; that is, for the editing states before long pauses (E-to-P), students' time spent on editing was memoryless. For the transitions between other states, the hazard rates are decreasing with time, indicating that the

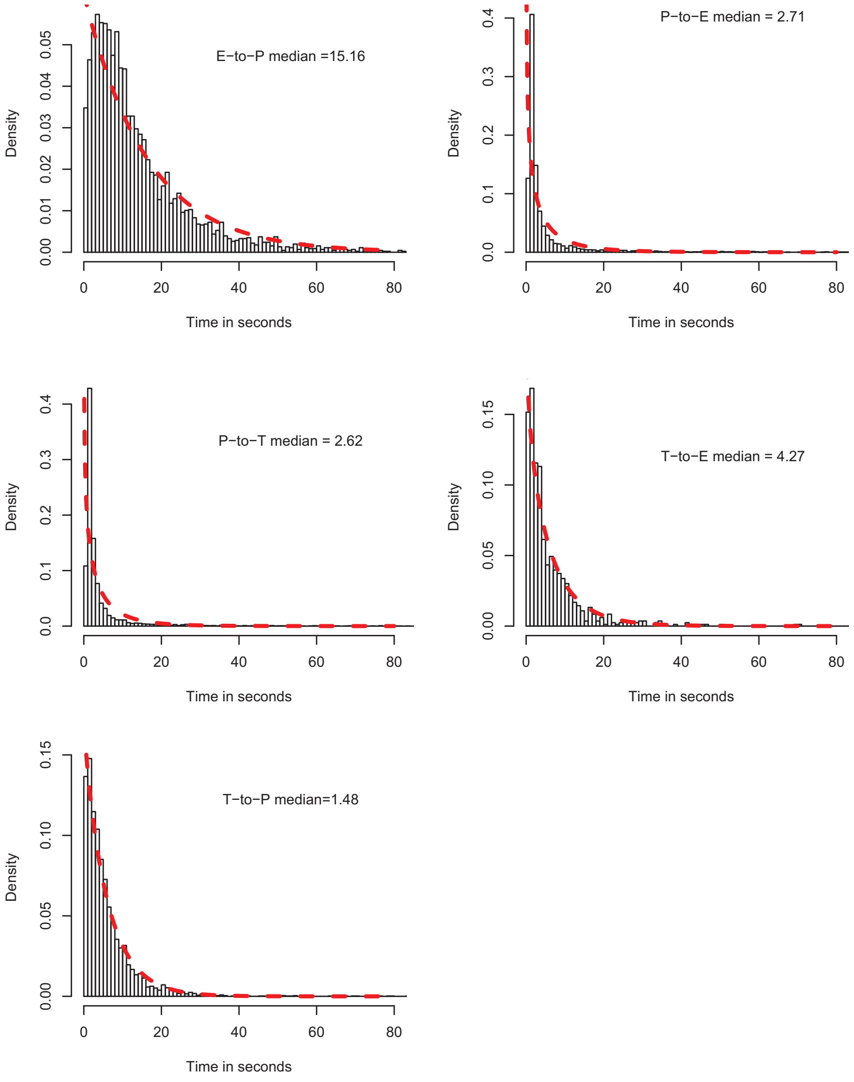

FIGURE 3. *Estimated density distributions of sojourn time in the semi-Markov model with $v_{12} \equiv 1$ compared to empirical data.*
Source. Copyright by Educational Testing Service, 2019. All rights reserved.

longer they stayed in a certain state, the less likely they were to leave that state; the sojourn time has a relatively large impact on the hazard rates within 50 seconds or so; after that, the sojourn time has less impact since the hazard function curves are flattened out and there were few observations.

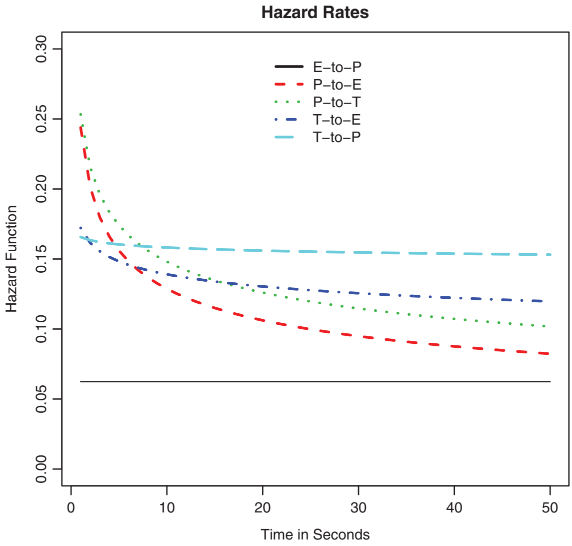

FIGURE 4. *Hazard rates of the sojourn times in the semi-Markov model with constraint* $v_{12} \equiv 1$.
Source. Copyright by Educational Testing Service, 2019. All rights reserved.

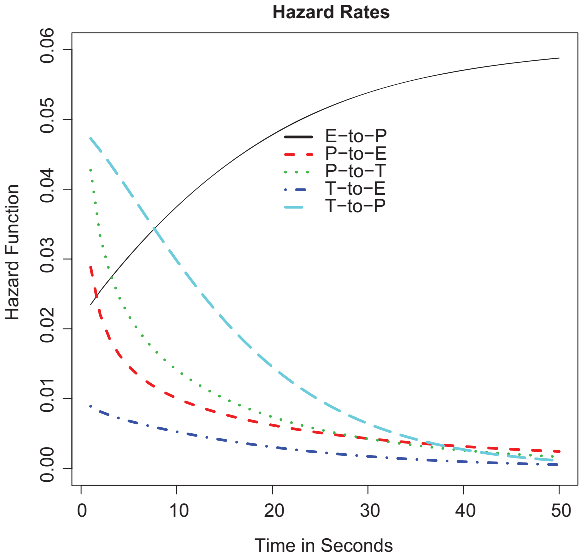

FIGURE 5. *Hazard rates of the transition probability in the semi-Markov model with* $v_{12} \equiv 1$.
Source. Copyright by Educational Testing Service, 2019. All rights reserved.

Figure 5 shows the estimated transition intensities (the hazard rate of the semi-Markov process) in the model with the constraint $\nu_{12} \equiv 1$. The curve for transition "E-to-P" suggests that, after 50 seconds or so, the students' instantaneous transition probability from E to P was relatively high and maintained at a constant rate up to 100 or 200 seconds (beyond 200 seconds, observations were sparse). For the remaining transitions, the instantaneous transition probabilities decreased quickly to 0, indicating negligible instantaneous transitions after 50 seconds or so.

Overall, from the above plots, we observed that the hazard functions were functions of time within a duration of 50 seconds or so. Beyond that, there were too few transitions to make meaningful claims.

For the following subgroup process comparisons, we used the full semi-Markov model without model selection (Berk, Brown, & Zhao, 2010). Similar general observations for the hazard rates were found for the subgroup analyses, which are not presented to avoid duplication. Instead, we focus on the subgroup differences from now on; that is, we focus on the beta parameters in the Cox regression model in Equation 1.

3.2. Subgroups

In each of the following analyses, a matching procedure (as described earlier) was conducted, so that the studied groups were equivalent in terms of their RS1 scores[2] and then the semi-Markov model was fit to the matched data with the covariate of the group variable (SES, race, or gender). In this way, we sought to obtain informative differences in writing processes between subgroups while conditioning on the same writing proficiency.

We present tables for β estimates to compare hazard rates of the sojourn time (A1). In those tables, when β was not significantly different from zero, the two studied groups were regarded equal in their sojourn time for the transition. Note that these hazard ratios of sojourn time are constants by the Cox model in Equation 1. However, the hazard ratios of the process, ratios of $\lambda(t)$ in (A2), are more complicated and have to be presented graphically.

In addition to the above results from semi-Markov models, for each subgroup, we provide summaries of the following variables as an aid to interpreting the results from the semi-Markov models. These variables are computed for each student:

- *Number of words*: the total word count in the essay.
- *Word length*: the median of the word lengths in the essay.
- *Standardized frequency index (SFI)*: SFI of the unique words (SFI is a measure of English vocabulary complexity. The higher the value, the more common the word; a lower value indicates more complex vocabulary; Zeno, Ivens, Koslin, & Zeno, 1995). The median of SFIs is used as a summary statistic.

TABLE 5.
Summary Statistics of Essay Production by Socioeconomic Status Groups

	n	Number of Words	Word Length	SFI	Writing Efficiency	Seconds per Word	Seconds per Character	RS1	Total Time
Mean (high)	92	165.30	3.90	66.52	.75	.81	.21	1.83	625.84
SD (high)		86.33	0.31	2.52	.07	.28	.08	0.81	385.42
Mean (low)	46	162.87	3.87	67.23	.71	.93	.24	1.72	621.18
SD (low)		87.78	0.32	1.99	.07	.42	.08	0.89	303.20

Note. The *t* tests of mean differences for writing efficiency and seconds per character are statistically significant at $p < .05$.

Source. Copyright by Educational Testing Service, 2019. All rights reserved.

- *Writing efficiency*: the ratio of the number of characters in the final essay to the total number of keystrokes produced in the process.
- *Seconds per word* and *seconds per character*: the mean duration of words and characters (i.e., the total time divided by the total number of words and characters), respectively.
- *Total time*: the total writing time from the first keystroke to the last one in seconds.

Seconds per words and seconds per character are measures of keyboarding skills. Word length and SFI are measures of vocabulary complexity. The first three features are word features that can be observed in the final essay, and the rest are keystroke features that cannot be derived from the final essay product.

Note that in the following tables, a statistical significance test was conducted separately for each target quantity without adjusting the significance level of the critical value (i.e., 5%).

3.2.1. SES. Table 5 shows the feature means and *SD*s for the SES groups. As expected, the two groups were quite close in RS1 writing scores as a result of the matching procedure (the second to the last column). The same holds true for the three word features: word count, word length, and SFI. However, given that they produced essays of equal quality in similar lengths of time, compared to the high SES group, the low SES group showed statistically lower writing efficiency and slower typing character-wise.

Table 6 shows the estimated βs for the SES variable in Equation 1 from the semi-Markov model. The statistic follows a χ^2 distribution of degree of freedom 1. From Table 6, the estimated coefficients $\beta_{32} = -.60$ and $\beta_{12} = -.20$ are statistically significantly negative for the low SES students, indicating that these students spent longer time at E and T before entering P, compared to the high SES students. The estimated coefficient $\beta_{23} = .12$ is statistically significantly positive, indicating that, when transitioning from P to T, the low SES students

TABLE 6.
Estimated βs for the Socioeconomic Status Groups in the Full Semi-Markov Model

Label	Transition	$\hat{\beta}$	*SD*	LCI	UCI	H_0	Statistic	*p* Value
β_{12}	E → P	−.20	.05	−.30	−.11	.00	17.01	<.0001
β_{21}	P → E	−.03	.05	−.12	.07	.00	0.32	.5716
β_{23}	P → T	.12	.04	.04	.20	.00	8.56	.0034
β_{31}	T → E	.14	.11	−.07	.35	.00	1.70	.1923
β_{32}	T → P	−.60	.04	−.67	−.53	.00	253.32	<.0001

Note. 1, 2, and 3 in the label column stand for E, P, and T, respectively. LCI = lower confidence interval; UCL = upper confidence interval; *SD* = standard deviation.
Source. Copyright by Educational Testing Service, 2019. All rights reserved.

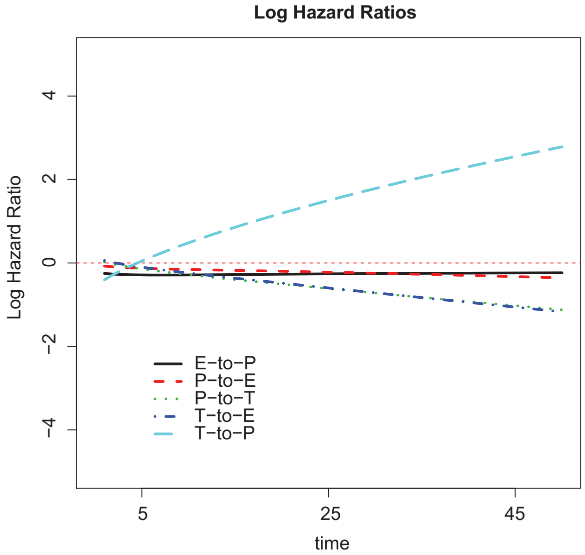

FIGURE 6. *Logarithm of the hazard ratios of transition intensities for the low socioeconomic status (SES) group and for the high SES group. The dashed horizontal line stands for a log ratio of zero (when the transition intensities for the two groups are equal).*
Source. Copyright by Educational Testing Service, 2019. All rights reserved.

spent shorter time at P before entering T. The estimated coefficient $\beta_{31} = .14$ is positive as well, though not statistically significantly, indicating that the low SES students in the sample spent shorter time at T before entering E.

Figure 6 shows the logarithm of the hazard ratios of transition intensity $\lambda_{ij}(t)$ (hazard rate of the process) for the two SES groups in the semi-Markov process.

TABLE 7.
Summary Statistics of Essay Production by Racial/ethnic Groups

	N	Number of Words	Word Length	SFI	Writing Efficiency	Seconds per Word	Seconds per Character	RS1	Total Time
Mean (White)	66	138.91	3.85	67.11	.74	.89	.23	1.42	540.46
SD (White)		71.25	0.35	2.51	.08	.42	.10	0.56	297.99
Mean (Black)	22	137.59	3.86	67.24	.70	.94	.24	1.41	626.21
SD (Black)		95.35	0.35	2.64	.06	.25	.06	0.59	474.00

Note. The *t* tests of mean score differences for writing efficiency are statistically significant at $p < .05$.
Source. Copyright by Educational Testing Service, 2019. All rights reserved.

TABLE 8.
Estimated βs for the Racial/Ethnic Groups in the Full Semi-Markov Model

Label	Transition	$\hat{\beta}$	*SD*	LCI	UCI	H_0	Statistic	*p* Value
β_{12}	E → P	−.42	.07	−.56	−0.29	.00	36.21	<.0001
β_{21}	P → E	.03	.07	−.10	.16	.00	0.24	.6242
β_{23}	P → T	.17	.06	.06	.28	.00	8.67	.0032
β_{31}	T → E	.48	.15	.19	.78	.00	10.22	.0014
β_{32}	T → P	−.56	.06	−.67	−.45	.00	98.33	.0001

Note. LCI = lower confidence interval; UCL = upper confidence interval; *SD* = standard deviation.
Source. Copyright by Educational Testing Service, 2019. All rights reserved.

Results indicate that the low SES group had a lower transition intensity for T-to-P within 5 seconds but had a much higher transition intensity than the high SES group after the 5-second duration. The other transition intensities of the low SES group were lower. That is, the low SES students only made more frequent transitions for T-to-P for longer sojourn times.

Overall, the above comparison results for the SES groups show that, to produce essays of the same quality as the high SES group, the low SES group spent significantly longer time at States T and E, which might be caused by their lower writing efficiency and slower character-level typing speed. In addition, they seemed to spend limited time in pauses/planning before producing text. Finally, they made less frequent transitions from state to state except for T-to-P, compared to the high SES group.

3.2.2. Race/ethnicity. Table 7 gives the summary statistics for the two racial ethnic groups examined, Blacks and Whites. Again in Table 7, the two groups

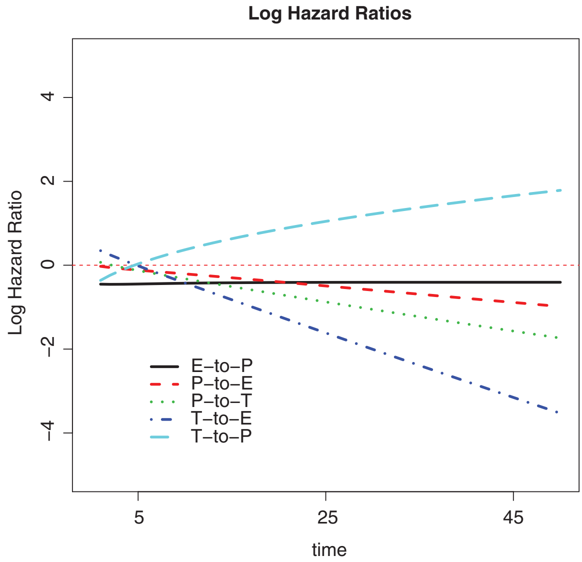

FIGURE 7. *Logarithm of the hazard ratios of transition intensities for the Black students and White students. The dashed horizontal line stands for a log(ratio) of zero (when the transition intensities for the two groups are equal).*
Source. Copyright by Educational Testing Service, 2019. All rights reserved.

were quite close in writing scores as expected from the matching procedure; they had similar word-related features as well (number of words, word length, and SFI). Given equal essay quality, however, the Black students had statistically significantly lower text efficiency than the White students but similar keyboarding skills. Although, on average, the difference in total writing time was not statistically significant, the Black students in our sample spent sizably longer time on essay writing (Ziliak & McCloskey, 2008).

Table 8 shows the estimated β parameters for the race covariate from the full semi-Markov model. From Table 8, the estimated coefficients $\beta_{12} = -.42$ and $\beta_{32} = -.56$ are statistically significantly negative for the Black students, indicating that Black students spent longer time at States E and T before entering State P. The coefficients $\beta_{23} = .17$ and $\beta_{31} = .48$ are statistically significantly positive, suggesting that Black students spent shorter time at P before T and at T before E, respectively, compared to the White students. These observations suggest that Black students planned more quickly (planning less) and quicker in moving from text production to editing (making corrections, etc.)

Figure 7 shows the logarithm of the hazard ratios of transition intensities for the racial/ethnic groups in the full semi-Markov model. These results indicate that the Black students had a lower transition intensity for E-to-P than the White

TABLE 9.
Summary Statistics of Essay Production by Gender Group

	n	Number of Words	Word Length	SFI	Writing Efficiency	Seconds per Word	Seconds per Character	RS1	Total Time
Mean (male)	101	184.96	3.88	66.75	.74	.85	.22	2.12	696.31
SD (male)		88.48	0.33	2.14	.07	.35	.08	0.83	352.58
Mean (female)	101	196.55	3.93	66.05	.76	.77	.20	2.24	660.94
SD (female)		92.04	0.27	2.60	.08	.24	.06	0.95	340.40

Note. The *t* tests of mean differences for SFI and second per character are statistically significant at $p < .05$. SFI = standardized frequency index.
Source. Copyright by Educational Testing Service, 2019. All rights reserved.

TABLE 10.
Estimated βs for the Gender Groups in the Full Semi-Markov Model

Label	Transition	$\hat{\beta}$	*SD*	LCI	UCI	H_0	Test	*p* Value
β_{12}	E → P	.29	.04	.21	.36	.00	55.39	<.0001
β_{21}	P → E	−.02	.03	−.09	.05	.00	0.39	.5323
β_{23}	P → T	.18	.03	.12	.24	.00	36.06	<.0001
β_{31}	T → E	.02	.08	−.13	.17	.00	0.06	.8065
β_{32}	T → P	−.57	.03	−.62	−.52	.00	479.02	<.0001

Note. 1, 2, and 3 in the label column stand for E, P, and T, respectively. LCI = lower confidence interval; UCL = upper confidence interval; *SD* = standard deviation.
Source. Copyright by Educational Testing Service, 2019. All rights reserved.

students; they also made less frequent transitions for T-to-P within 5 seconds or so but a much higher transition intensity than the White students after that time. For the remaining transitions, the Black students were more likely or equally likely to make change states within the 5-second duration but less likely to do so after 5 seconds, compared to their counterpart students.

Again, from the above results, to produce essays of the same quality as the White students, the Black students spent statistically significantly longer time in the text production and editing states before long pauses and tended to make quicker and shorter transitions in getting to those states.

3.2.3. Gender. Table 9 shows the feature summary statistics for the gender groups. Again, the two groups were quite close in writing scores due to the matching procedure. Given that they produced essays of equal quality, the female group used statistically significantly more complex vocabulary (lower SFI) and

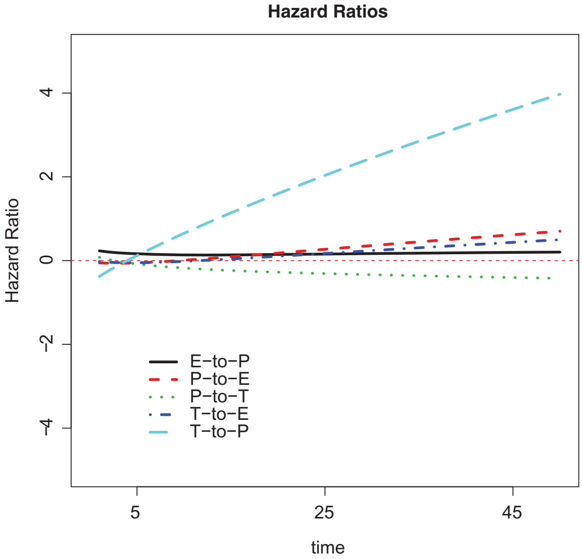

FIGURE 8. *Logarithm of the hazard ratios of transition intensities for the gender groups. The dashed horizontal line stands for a log(ratio) of zero (when the transition intensities for the two groups are equal).*
Source. Copyright by Educational Testing Service, 2019. All rights reserved.

was significantly faster in typing on the character level. In addition, female students in our sample appeared to have produced essays of about 10 words longer and used shorter total time in writing compared to the male students; these differences were not statistically significant but noticeable (Ziliak & McCloskey, 2008).

Table 10 shows the estimated β parameters for the gender groups from the full semi-Markov model. The estimated $\beta_{32} = -.57$ is statistically significantly negative, indicating the female students spent longer time at T before transitioning to P. The estimated coefficients $\beta_{23} = .18$ is statistically significantly positive, indicating female students spent shorter time at P before entering T compared to the male group. Notably different from the previous SES and racial/ethnic comparisons, the estimated coefficients $\beta_{12} = .29$ is also statistically significantly positive, indicating female students spent shorter time at E before entering P.

Results of the logarithm of the hazard ratios in Figure 8 revealed that, compared to the male group, the female group made more frequent E-to-P transition across the time span. For P-to-E and T-to-E, they made similar or more frequent transitions; for P-to-T, they made similar or less frequent transitions. The female group made less frequent T-to-P transitions within the 5-second duration, but they were more likely to make T-to-P transitions after that.

Overall, from the above gender comparison, we may infer that the female group was more fluent in writing than the male group even after score matching: On average, female students were faster in character-level typing, they used more complex words, they spent longer time in continuous text production before making long pauses, and they were more likely to make short and quick editing and pause transitions. These characteristics may explain why the female students in our sample produced relatively long essays (10 words longer) in shorter total writing time (40 seconds shorter).

4. Discussion

In this study, we used keystroke logs to model the writing processes and investigated differences in writing processes between subgroups given the same essay quality. We focused on the duration time that students spent in each state (long pauses, editing, or text production state) and the probabilities of transitions from one state to an other. We found that differences in writing processes existed between subgroups given the same quality of essays, as well as across the three demographic groups studied. For example, both the low SES students and the Black students showed significantly lower efficiency in text production than did comparison groups; that is, their final texts were smaller portions of the total keystroke events compared to the higher SES students and White students, respectively. In contrast, compared to male students, the female students were more fluent, typing faster, using more complex words, spending longer time in text production, and engaging in quick and frequent editing and pauses.

While there appear to be notable writing process differences across the three demographic comparisons, a common observation is that the studied focal groups (low SES students, Black students, and female students) spent longer time at the text production states before entering the long pause states, and then they made less frequent transitions within 5 seconds or so for T-to-P. However, the meaning of this finding might not necessarily be the same for each group. For example, the low SES students and Black students might have spent more time in text production because they appeared to need to expend greater effort to produce essays of the same quality as their counterparts (indicated by their lower text efficiency). For the female students, the additional time in text production was also associated with greater fluency.

These results are consistent with what the literature suggests for the potential causes of lower writing performance. For instance, difficulties with transcription skills (keyboarding and spelling) can reduce the amount of working memory available for other writing processes such as planning and evaluation. This reduction may lead to a more serial writing strategy. In such a strategy, students with less fluent transcription skills may need to spend a greater amount of time producing and editing the text (e.g., as in Table 6, the low SES group had significantly negative β parameters β_{32} and β_{12}), switching between states less often (e.g., as in Figure 6, the hazard rations are below 1), since they may be unable to carry out evaluation or planning activities more concurrently with typing. Conversely, if students have strong transcription skills, they may be more likely to monitor their writing as they type and thus may be more likely to notice typographic errors, misspellings, or problematic phrasings quickly. This real-time monitoring ability might, in turn, make for quicker and shorter pauses and edits (as in Table 10, the female group had significantly positive beta parameters β_{23} and β_{12}). These are essentially compensatory processes. In any complex task, it may be possible to minimize the consequences of one skill being weak by devoting more time and effort or by taking advantage of areas of specific strength. Use of compensatory strategies would explain why some students from a disadvantaged group could achieve the same levels of performance as students from an advantaged group but display significant differences in their writing processes.

This study had several limitations including that we used a data set with small sample sizes. A larger sample size is preferable for modeling processes and for reducing random effects associated with individual students' responding. In addition, we used the propensity matching method to preprocess the data before comparing the subgroups, which may reduce information. Other statistical approaches that do not sacrifice data, such as supplying both a subgroup indicator and the matching variable as covariates in the Cox regression step, or supplying weights to standardize groups, are worth investigation in future analysis. These weighting methods include weighting by the minimum discriminant information and weighting adjusted for errorprone covariates (refer to Haberman, 2015, and Lockwood & McCaffrey, 2016).

Because of the specific writing task and limited sample sizes, the results obtained here may not be generalizable to a broader population or to different writing conditions, and the causal explanations we have suggested may or may not ultimately be confirmed. Further research might attempt to cross-validate our keystroke log-based writing state classification results in order to disambiguate what students are doing in each state, particularly during long pauses.

One additional limitation of this study is the time homogeneity assumption in the semi-Markov models. Common writing strategies such as whether writers edit more at a later stage of the writing session may call for heterogeneous models, which naturally is the next step in semi-Markov modeling. However, writing strategies common in the general population may not be associated with a particular subgroup leading to conclusions similar to those presented here. A final note is that statistical significance is not the same as practical significance; β estimates do not offer meaningful interpretation in terms of effect size (Ziliak & McCloskey, 2008) of subgroup differences.

Despite these limitations, this study offers important results. Perhaps the most important of those results is the potential value for analyzing writing process data of semi-Markov models over the more commonly used Markov models. Because they consider the continuous nature of writing state duration and by relaxing the memoryless property of the duration time distribution, semi-Markov models are well suited to the evaluation of such processes and group differences in them. Besides the studied continuous time stochastic processes, other alternative modeling approaches are worth exploring, including the dynamic Bayesian networks, if writing time is sliced into equal intervals (Mislevy, Almond, Yan, & Steinberg, 1999; Murphy, 2002). A second important result is that this study provides a supplement to existing sources of evidence about group differences in writing, which have (for the most part) focused on differences in scores and their correlations with socioeconomic variables. With inclusion of process differences, we add critical evidence that might help us to validate a more detailed causal account.

Appendix

The Semi-Markov Process

To introduce the semi-Markov models, let us consider a Markov renewal process (J_n, T_n), where $0 = T_0 < T_1 \ldots < T_n < \infty$ are the successive times of entry to states $J_0, J_1, \ldots, J_n$ where $J_n = J_{n+1}$ for all n. The sequence (J_n) is an embedded homogeneous Markov chain with transition probabilities $p_{ij} = P\,(J_{n+1} = j|J_n = i)$. Let $S_n = T_n - T_{n-1}$ be the duration time, and then the Markov renewal process satisfies

$$Q_{ij}(t) = P(J_{n+1} = j, S_{n+1} \leq t|J_0, \ldots, J_n = i, S_1, \ldots, S_n)$$

$$= P\,(J_{n+1} = j, S_{n+1} \leq t|J_n = i).$$

Let $N(t) = \sup\,\{n \in N : T_n \leq t, t \in R\}$ be the counting process that counts the total number of observed transitions during the time interval $[0, t]$. Then, $J_{N(t)}$ defines a homogeneous semi-Markov process; that is, the probability of a future state depends on both the current state and its sojourn time (Medhi, 1982).

Hazard Rates

Two hazard rates are introduced in this section: one is the hazard rate of sojourn time and the other is the hazard rate of the semi-Markov process.

Let $F_{ij}(t)$ be the probability distribution function of the sojourn time (staying at state i before entering state j), that is,

$$F_{ij}(t) = P(S_{n+1} \leq t|J_n = i, J_{n+1} = j) = \frac{Q_{ij}(t)}{p_{ij}}.$$

Let the survival function be $G_{ij}(t) = 1 - F_{ij}(t)$. *The hazard rate of the sojourn time* is defined as

$$\alpha_{ij}(t) = \frac{f_{ij}(t)}{G_{ij}(t)}, \tag{A1}$$

where $f_{ij}(t)$ is the density function. This hazard rate describes the likelihood of leaving the current state for the next one (Kalbfleisch & Prentice, 1980).

When the distribution of the sojourn time $F_{ij}(t)$ follows the exponential distribution that has one parameter $1/\sigma_{ij}$, the semi-Markov process is a continuous-time Markov chain mode, the sojourn time of which has the memoryless property; and in this case, the hazard rate of the sojourn time $\alpha_{ij}(t) = 1/\sigma_{ij}$ is constant (where σ corresponding to mean of the sojourn time and its *SD* at the current state): A larger hazard rate of α_{ij} implies a shorter average stay at the current state (*i*) before entering the next state (*j*). When the survival function follows the Weibull distribution that has a shape parameter ν and a scale parameter σ, the hazard rate $\alpha_{ij}(t) = \nu_{ij}/\sigma_{ij} \times (t/\sigma_{ij})^{\nu_{ij}-1}$ is a function of time *t*: Generally speaking, a larger hazard rate at time *t* indicates a shorter average sojourn time as well (when $\nu = 1$, the Weibull distribution degenerates to the memoryless exponential distribution; when $\nu < 1$, the hazard rate function decreases in the sojourn time; and when $\nu > 1$, the hazard rate function increases with the sojourn time; Krol & Saint-Pierre, 2015).

The hazard rate of the semi-Markov process (i.e., transition intensity) $\lambda_{ij}(t)$ corresponds to the probability of transition from state *i* to state *j* between time *t* and $t + \Delta t$, which is defined as

$$\lambda_{ij}(t) = \lim_{\Delta t \to 0} \frac{P(J_{n+1} = j, t < S_{n+1} \leq t + \Delta t | J_n = i, S_{n+1} > t)}{\Delta t},$$

$$= p_{ij}\alpha_{ij}(t)G_{ij}(t)(G_i(t))^{-1}, \tag{A2}$$

where $G_i(t) = \sum_{j=i} G_{ij}(t)$. The hazard rate $\lambda_{ij}(t)$ describes the instantaneous transition probability rate at time *t* from state *i* to state *j*. Given state *i* at time *t*, a larger hazard rate of $\lambda_{ij}(t)$ indicates a higher probability of entering state *j* (Medhi, 1982).

Authors' Note

Any opinions expressed in this publication are those of the authors and not necessarily of Educational Testing Service.

Acknowledgments

The authors would like to thank Daniel McCaffrey, Shelby Haberman, and especially two unanimous reviewers for their thoughtful comments on a previous draft of this article.

Declaration of Conflicting Interests

The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: The authors prepared the work as employees of Educational Testing Service.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. Because of very small sample sizes, other racial/ethnic groups were not studied.
2. The association of scoring rubric 2 (RS2) with keystroke features is weaker than that of RS1 (Guo, Deane, van Rijn, Zhang, & Bennett, 2018), so we only used RS1 in the matching in order to preserve more data. Other covariates were not used due to the same reason.

References

Allen, L., Jacovina, M., Dascalu, M., Roscoe, R., Kent, K., Likens, A., & McNamara, D. (2016). Entering the time series space: Uncovering the writing process through keystroke analysis. In T. Barnes, M. Chi, & M. Feng (Eds.), *Proceedings of the 9th International Conference on Educational Data Mining* (pp. 22–29). Raleigh, NC: International Educational Data Mining Society.

Baaijen, V., Galbraith, D., & de Glopper, K. (2012). Keystroke analysis: Reflections on procedures and measures. *Written Communication*, *19*, 246–277. doi:10.1177/0741088312451108

Berk, R., Brown, L., & Zhao, L. (2010). Statistical inference after model selection. *Journal of Quantitative Criminology*, *26*, 217–236.

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. doi:10.1207/s1532690xci1202

Cox, D. R. (1972). Regression models and life-tables. *Journal of Royal Statistic Society B*, *34*, 187–220.

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.

Guo, H., Deane, P., van Rijn, P., Zhang, M., & Bennett, R. (2018). Exploring the heavy-tailed features of keystroke logs in writing processes. *Journal of Educational Measurement*, *55*, 194–216.

Haberman, S. (2015). Pseudo-equivalent groups and linking. *Journal of Educational and Behavioral Statistics*, *40*, 254–273. doi:10.3102/1076998615574772

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

Ho, D., Imai, K., King, G., & Stuart, E. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, *15*, 199–236. doi:10.1093/pan/mpl013

Jackson, C. (2011). Multi-state models for panel data: the msm package for R. *Journal of Statistical Software*, *38*, 1–28. doi:10.18637/jss.v038.i08.

Kalbfleisch, J., & Prentice, R. (1980). *The statistical analysis of failure time data*. New York, NY: Wiley.

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

Kellogg, R. (2001). Competition for working memory among writing processes. *The American Journal of Psychology*, *114*, 175–191. doi:10.2307/1423513

Krol, A., & Saint-Pierre, P. (2015). *SemiMarkov*: An R package for parametric estimation in multi-state semi-Markov models. *Journal of Statistical Software*, *66*, 1–16. doi:10.18637/jss.v066.i06

Leijten, M., & van Waes, L. (2013). Keystroke logging in writing research: Using input log to analyze and visualize writing processes. *Written Communication*, *30*, 358–392.

Lockwood, J. R., & McCaffrey, D. F. (2016). Matching and wighting with functions of error-prone covariates for causal inference. *Journal of American Statistical Association*, *111*, 1831–1839. doi:10.1080/01621459.2015.1122601

Matsuhashi, A. (1981). Pausing and planning: The tempo of written discourse production. *Research in the Teaching of English*, *15*, 113–134.

McCutchen, D. (1996). A capacity theory of writing: Working memory in composition. *Educational Psychology Review*, *8*, 299–325. doi:10.1007/BF01464076

Medhi, J. (1982). *Stochastic processes*. New York, NY: Wiley. ISBN 978-0-470-27000-4.

Mislevy, R. J., Almond, R. G., Yan, D., & Steinberg, L. S. (1999). Bayes nets in educational assessment: Where do the numbers come from? In K. B. Laskey & H. Prade (Eds.), *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence* (pp. 437–446). San Francisco, CA: Morgan Kaufmann.

Murphy, K. (2002). *Dynamic Bayesian networks: Representation, inference and learning*. Berkeley, CA: UC Berkeley, Computer Science Division.

National Center for Education Statistics. (2011). *Writing 2011: The National Assessment of Educational Progress at grades 8 and 11* (NCES-2012-470). Washington, DC: Institute for Education Sciences, U.S. Department of Education.

R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/.

Rosenbaum, P., & Rubin, D. (1983). The central role of the propensity score in observational studies for causal effecs. *Biometrika*, *70*, 41–55.

Schilperoord, J. (2002). On the cognitive status of pauses in discourse production. In T. Olive & M. C. Levy (Eds.), *Contemporary tools and techniques for studying writing* (pp. 61–87). Dordrecht, the Netherlands: Kluwer Academic. doi:10.1007/978-94-010-0468-8

Zeno, S., Ivens, S. M., Koslin, S. H., & Zeno, B. L. (1995). *The educator's word frequency guide*. Brewster, NY: Touchstone Applied Science Associates.

Zhang, M., Bennett, E. R., Deane, P., & van Rijn, P. (2019). Are there gender differences in how students write their essays? An analysis of writing processes. *Educational Measurement: Issues and Practice*. doi:10.1111/emip.12249.

Zhang, M., & Deane, P. (2015). *Process features in writing: internal structure and incremental value over product features* (ETS RR-15-27). Princeton, NJ: Educational Testing Service.

Zhang, M., van Rijn, P., Deane, P., & Bennett, E. R. (2019). Scenario-based assessments in writing: An experimental study. *Educational Measurement: Issues and Practice*. doi:10.1080/10627197.2018.1557515

Ziliak, S., & McCloskey, D. (2008). *The cult of statistical significance: How the standard error costs us jobs, justice, and lives*. Ann Arbor: University of Michigan Press. doi:10.3998/mpub.186351

Authors

HONGWEN GUO is a senior research scientist at Educational Testing Service, 660 Rosedale Rd., MS-12T, Princeton, NJ 08541; email: hguo@ets.org. Her primary research interests are psychometric and statistical modeling and analyses.

MO ZHANG is a research scientist at Educational Testing Service, 660 Rosedale Rd., MS-03T, Princeton, NJ 08541; email: mzhang@ets.org. Her work has centered on automated scoring of constructed-response items, timing and process analyses for writing, and performance-based assessment design and analyses.

PAUL DEANE is a principal research scientist at Educational Testing Service, 660 Rosedale Rd., MS-11R, Princeton, NJ 08541; email: pdeane@ets.org. His research interests center on reading, writing, and vocabulary learning and assessment.

RANDY E. BENNETT is Norman O. Frederiksen Chair in Assessment Innovation at Educational Testing Service, 660 Rosedale Rd., MS-02R, Princeton, NJ 08541; email: rbennett@ets.org. His research interests center on integrating advances in cognitive and learning science, measurement, and technology to devise assessment approaches that have positive impact on teaching and learning.

Manuscript received May 31, 2018
Revision received October 23, 2018
Accepted April 23, 2019