Effects of Scenario-Based Assessment on Students' Writing Processes

Hongwen Guo
Educational Testing Service
hguo@ets.org

Mo Zhang
Educational Testing Service
mzhang@ets.org

Paul Deane
Educational Testing Service
pdeane@ets.org

Randy E. Bennett
Educational Testing Service
rbennett@ets.org

This study investigates the effects of a scenario-based assessment design on students' writing processes. An experimental data set consisting of four design conditions was used in which the number of scenarios (one or two) and the placement of the essay task with respect to the lead-in tasks (first vs. last) were varied. Students' writing processes on the essay task were recorded using keystroke logs. Each keystroke action was classified into one of four writing states: planning, text production, local edit, or jump edit, and a semi-Markov model was fit to the data. Results showed that the single-scenario and essay-last design encouraged fewer but longer editing states compared to the alternative designs. Additionally, this task ordering appeared to have enabled more fluent and efficient text production when paired with a single scenario. These results seem explainable from cognitive writing theory, particularly with respect to working memory load. Limitations and future directions for research are also discussed.

Keywords: semi-Markov process, keystroke logs, writing instruction

1. INTRODUCTION

In advanced academic and career environments, individuals are routinely expected to know how to draw on information from multiple sources and integrate them into a coherent written argument. As a result, writing from sources is a competency frequently contained in English language arts content standards (e.g., Common Core State Standards Initiative 2010, p. 18) and a task type found on many state accountability assessments, including those offered by the Smarter Balanced Assessment Consortium (SBAC) and the Partnership for Assessment of Readiness for College and Careers (PARCC). One approach to assessing writing from sources is scenario-based assessment (Deane et al., 2011). In this theoretically driven approach, a scenario (or topical context) is presented along with source materials. Students are then given a sequence of lead-in tasks that requires reading and summarizing arguments in the sources, critiquing those arguments, analyzing them, and finally composing an essay presenting a position and reasoning using evidence from the sources. The scenario is included to increase engagement by providing a reasonably realistic setting; the lead-in tasks facilitate engagement with the sources, reducing differences in topic familiarity, and activating skills needed for completing the essay task

(Bennett et al., 2016; Deane et al., 2011). This structure simulates a condensed writing project undertaken in an order that a skilled practitioner might follow.

A key question with respect to this assessment design is how well it functions psychometrically, both at the test level and more particularly for the essay task's measurement of argumentative writing skill. Because the standardized assessment of writing skill has traditionally been done without lead-in questions, it is reasonable to ask whether the scenario-based structure has a facilitating or degrading impact on student performance and on score validity.

To address this question, Zhang et al. (2019) conducted an experiment in which they varied both the position of the essay and whether the essay and lead-in tasks each drew upon the same or on different scenarios and source materials. They found that, with a single scenario, presenting the essay before or after the lead-in questions had no effect on either mean essay score or mean total test score. However, students in the condition of the original order of tasks (i.e., single scenario & essay-last) spent less time writing and produced shorter essays of similar quality, suggesting that the design improved efficiency. When the authors compared the scenario-based design to the mixed-scenario conditions, they found that a mixed scenario with the essay last resulted in fewer words written, less time on task, and lower scores, suggesting that this design may have reduced motivation and engagement. While the essay-first mixed-scenario condition produced essay scores similar to the scenario-based design, students in the mixed condition used more words and had to spend more total time writing to achieve those similar scores.

In a complementary analysis of the data with only those students taking the single-scenario conditions, Zhang et al. (2018) regressed essay scores on process measures. They found that when the essay followed the lead-in questions, the essay scores were considerably less dependent on fluency-process features and more related to local-editing features (e.g., extent of typo and minor word-correction). These results suggest that the preparation afforded by the lead-in tasks reduced the effect of fluency, allowing students extra time to become familiar with the topic, organize their position and, consequently, more efficiently and presentably render it. Doing well on the essay-first condition, in contrast, seemed to favor students adept at rapid idea generation, expression, and text entry.

The current study delves more deeply into how the scenario-based structure impacts students' writing processes. In particular, we compare design conditions in terms of the *writing states* that students evidence in composing their essays and the time they spend in those states. A common decomposition of the writing process identifies four states: planning, translation, transcription, and revision (Hayes, 2012; Kellogg, 2001). Planning involves task analysis, idea generation and text organization. Translation includes the linguistic operations needed to express ideas in words and sentences. Transcription denotes the process of rendering that language on paper or on screen. Finally, revision involves reviewing and amending the text to correct errors or to otherwise improve the text content or the plan underlying it.

In the cognitive writing research literature, investigators have used the keypress, mouse click, and latency information from keystroke logs to infer the presence of these states to better understand writing processes (Baaijen et al., 2012). These methods have also been applied to understanding composition processes in the context of writing assessment (Deane et al., 2011; Deane and Zhang, 2015; Deane et al., 2019; Guo et al., 2018; Leijten and Van Waes, 2013; Zhang et al., 2018; Zhang et al., 2019). Most relevant to the current study, Guo et al. (2019) used a three-state Markov-type model to study students' writing processes dynamically, finding differences between student subgroups matched on essay score.

The current study uses an elaboration of the Guo et al. (2019) method to better understand

the differences among the test designs in terms of writing processes, thereby complementing the original Zhang et al. (2019). A fuller range of writing states was operationalized in line with cognitive writing theory by incorporating a new jump-edit state to represent a more sophisticated self-monitoring process during writing (in contrast to making small and localized fixes usually on the word level). In particular, we used keystroke logs to automatically classify each of the actions occurring over the course of composition into the following four states:

- A Long Pause state (P) generally associated with planning. Long pauses are normally located at the beginning of the writing process or at natural linguistic and sentence boundaries.
- A Text Production state (T) that combines the translation and transcription process. This state represents text generation –i.e., translating ideas into language and transcribing them into written text. A typical T state is composed of keystroke actions implemented fluently and consecutively in the form of bursts of text generation without editing or interruption by a long pause.
- A Local Editing state (E) normally indicated by small fixes or typo-corrections to words or phrases. This state may suggest the extent of self-monitoring occurring as part of the revision process.
- A Jump Editing state (J) characterized as a move to a new location, usually further away from the current word, in order to make changes to the text. The intention was to distinguish jump behaviors in the writing process from more localized editing activities. Jump behaviors imply the occurrence of global reviewing and revision, including possibly making substantive changes to the text content.

The definition of this second revision state, J, can be viewed as a theoretically significant addition because jump-edit is a meaningful behavior distinct from other text-production and revision activities (Deane et al., 2019; Zhu et al., 2019). This proposition is investigated empirically through stochastic modeling in the current study.

The remainder of the paper is structured as follows. First, we describe the writing assessment, the test design variations, the data collection, and basic summaries of scores and response time. We next describe the classification procedure of keystroke log data for writing states, and then briefly discuss the semi-Markov models and their interpretation. Third, we present the results of model fitting, and then fit semi-Markov models to the data so that the four writing states and their duration times can be compared among different assessment designs. Finally, we summarize the study and discuss the implications of our results.

2. Data

2.1. Test Forms

This study used data collected in a prior study by Zhang et al. (2019). That study employed a scenario-based English language arts (ELA) summative writing assessment developed as part of the CBAL® research program at ETS (Bennett et al., 2016; Deane et al., 2011). The topic, or scenario, used in this test form concerns whether the US should ban advertising to children. This topic is hereafter designated as BA. Three reading sources were given, with each source

presenting a particular point of view. The source materials, which were available throughout the assessment, included adaptations of magazine or news articles (mean number of words per source was 246) appropriate for middle schools.

The original (single-scenario & essay last) assessment form and three variations of it were used in experimental fashion to evaluate the design effects. In the original form condition, an essay task (Designed Task 4) followed three lead-in tasks (Tasks 1, 2, and 3), with the order of the tasks intended to help engage students with the sources and activate critical thinking skills important for writing the essay. This order recapitulated the order that an experienced writer might follow in composing an argumentative essay (Bennett et al., 2016; Deane et al., 2019). This form is hereafter referred to as Form 1. Forms 2, 3, and 4 were built to give variations on Form 1's scenario structure (e.g., by changing the task order or by including a second scenario).

Table 1 shows the organization of each form into two sections administered in separate sessions. Form 2 had the same topical setting, purpose, and source materials as Form 1 (both are single-scenario-based forms). However, in Form 2, the essay (with the source materials) was given before the lead-in questions. Form 3 employed the same ordering as Form 1 but with a different setting and source materials for the lead-in tasks from that of the BA essay; in this case the lead-in tasks are about whether students should be paid cash for receiving good grades in school (designated CG). Like Form 1, the three sources were adapted from magazine or news articles (278 words on average). The essay task, however, used the same topical setting, purpose, and source materials as in Form 1. That is, Form 3 had a mixed-scenario-based form. Form 4 also had a mixed scenario, but with the BA essay presented first followed by the same three CG lead-in tasks as in Form 3.

Table 1 also shows that the three lead-in tasks (Tasks 1, 2, and 3) have 9, 1, and 7 items respectively. In terms of item format and relative position, the lead-in tasks were the same across forms. Their content, however, was aligned to the scenario with which they were associated (i.e., BA or CG). Lead-in Task 1 and Task 2 were taken in the same session, and Task 3 and Task 4 (BA Essay) were administered in another session. Each student took the two sessions, each session had a 45-minute time limit, and students were not expected to leave until test completion.

Table 1: Four test forms, organization of testing sessions, and task definitions

Form	Session 1		Session 2	
1 (Single-scenario&essay-last)	BA Task 1	BA Task 2	BA Task 3	BA Essay
2 (Single-scenario&essay-first)	BA Essay	BA Task 3	BA Task 1	BA Task 2
3 (mixed-scenario&essay-last)	CG Task 1	CG Task 2	CG Task 3	BA Essay
4 (mixed-scenario&essay-first)	BA Essay	CG Task 3	CG Task 1	CG Task 2
	Task Information			
Task 1	Read and summarize arguments (30 mins)			9 items
Task 2	Evaluate the argument in a letter to the editor (15 mins)			1 item
Task 3	Analyze arguments (10 mins)			7 items
BA Essay	Present your view in an essay (35 mins)			1 prompt

Note. BA = Whether the US should ban advertising to children; CG = Whether students should be paid cash for receiving good grades in school.

The essays were graded by human raters against two rubrics, each on an integer scale running from 0 to 5[1]. The writing fundamentals rubric (denoted as RS1) evaluated such basic writing aspects as grammatical correctness, word usage, mechanics, organization, and development, whereas the other rubric (denoted as RS2) evaluated such higher-level skills as the strength of the evidence and quality of arguments. For each rubric, scoring was done holistically; that is, for each score level the rubric described the characteristics that a response at that level should possess. Raters used those descriptions, explanatory notes, and example student responses to give a score for an essay (see Zhang et al. 2019 for detailed information about rater training and rater agreement, and see Deane and Zhang 2015, Zhang et al. 2019, and Zhang et al. 2019 for a discussion on writing features).

2.2. Student Group Information

Data were collected from 1,082 students attending the 8th grade from eight volunteer schools in New Jersey, Delaware, West Virginia, South Carolina, Alabama, Minnesota, South Dakota, and Utah over a one-month period between September and October 2014. Within each classroom, students were assigned at random to one of the four test forms. In the later analysis, Group 1 refers to students who took Form 1, Group 2 refers to those who took Form 2, and Group 3 and Group 4 are defined similarly. Because of the random design, the four student groups are very similar in terms of their writing skills. We also considered Group24, which is the combination of students in Group 2 and Group 4 who received the same writing prompt at the very beginning of the test.

Besides the assessment, teacher ratings of students' argumentative writing skills were collected. Ratings were rendered on a 1-5 scale, with 5 indicating the most advanced level.

3. Method

3.1. Classification of Writing States

To classify the sequence of actions into different writing states for each student, our classification algorithm conducted the following steps.

1. Defined the gap (or pause) time between two adjacent keystroke actions (interkey interval; IKI), specifically the pause time that preceded an action.

2. Obtained the personalized in-word typing speed index: the keyboarding skill (KBS). KBS was defined for each writer as his or her within-word median time (in seconds) per character for typing words taken from the Oxford English Corpus list of 100 most frequently used words and their inflections.

3. Used KBS $\times L$ as the threshold for defining a long pause for each individual student, where L is a pre-specified number. In our study, L=10 was chosen, so that ten times KBS approximates the time required to type two commonly used words for that individual

[1]For the purposes of our analyses, we excluded responses receiving a human score of 0 (a few students on each form) that indicated empty or off-topic responses, plagiarized responses, and responses consisting of random keystrokes.

student. Such a personalized threshold was larger than the 95th percentile of within-word pauses for the most frequently used 100 words for all students in the studied sample[2].

4. Classified an action as a P state if the action was associated with a long pause (i.e., IKI > threshold). The action sequences in between P states were considered as non-P chunks.

5. For a non-P chunk, if it did not contain any cut or paste actions or more than two deletes or replaces or combined actions, classified the chunk as a T state; otherwise the chunk was temporarily given a U (undecided) state.

6. For a U chunk, used a sliding window to scan across the chunk to decide whether the actions in a window were in an E or T state. Within each window of a pre-specified size of W[3], if more than two deletes or replaces or combined actions were detected, the actions in that window were assigned an E state. Otherwise, the actions in the window were assigned to a T state.

   After this step, every action was assigned a state E, P, or T.

7. Independently, identified jump back (JB) and jump forward (JF) actions using the position of an action in the text. Defined a jump-editing state (J) for a cycle starting from a JB and ending with a JF[4]. All the actions in the cycle were labeled as a J state.

   Updated the state assignment for all actions, with a J state always overwriting any of the E, P, or T states.

8. Finally, treated the consecutive same states as one state and aggregated the corresponding IKIs, so that a state did not transition to itself.

Figure 1 shows an example of the state sequence obtained from one student's keystroke log. The x-axis is the time in seconds recorded in his writing process, and the y-axis is the identified writing state. In this example, the student started with a long P state (probably planning what to write), and moved through E, J, P, T states with various duration times, and ended in a J state (likely a global editing behavior). This student had few E and J states, and many P and T states during the writing session, which he finished in about 30 minutes. His total number of states is 160, with a RS1 score of 1.

3.2. Semi-Markov Processes

In this study, we used Markov-type processes because they are suited for modeling discrete states and continuous duration times such as those found in composition (refer to Figure 1). Based on the results of previous research on stochastic modeling of writing processes, we focused on semi-Markov models (Guo et al., 2019; Krol and Saint-Pierre, 2015). A semi-Markov model contains observations of states $\{S_i, i = 1, 2, \cdots, I\}$ and their duration times $\{T_{S_i,S_j}\}$ at state S_i before

[2]A personalized threshold was necessary, rather than a single fixed threshold for all students, because students vary widely in their typing facility. Because they laboriously enter text letter by letter, poor typists may have many long pauses between keystrokes that are not indicative of planning behavior. For these students, capturing likely planning behavior requires a longer pause threshold that takes account of their typical typing speed. A more complete justification for this measure can be found in Zhang et al. (2019) and Zhang et al. (2018).

[3]W=5 was chosen in our study as the average word length.

[4]When there were multiple JF actions in the cycle, we used the last JF to close the cycle.

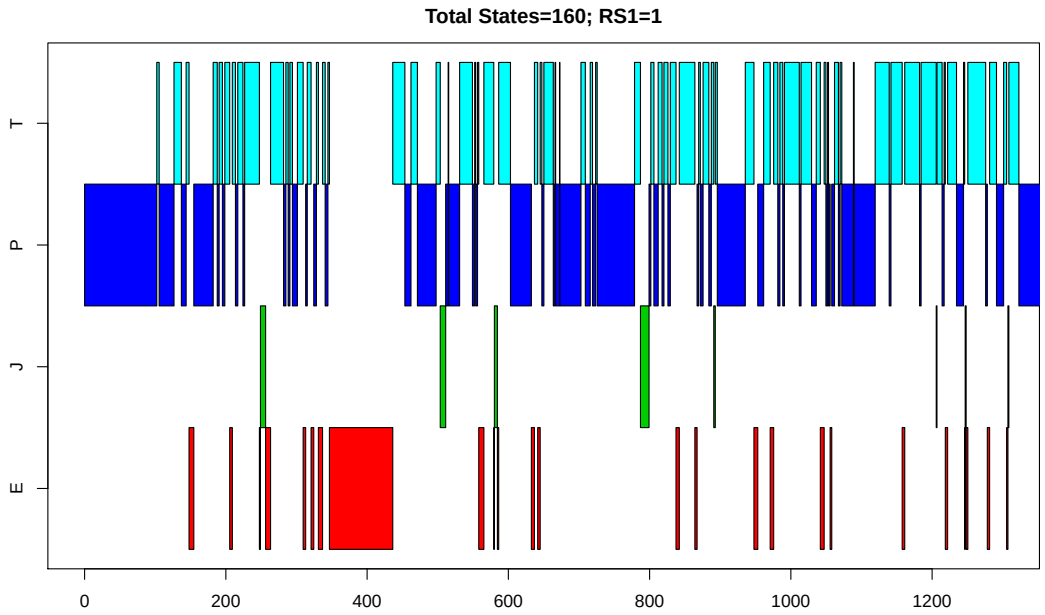

Figure 1: The state sequence of one student. The x-axis stands for time in seconds, and the y-axis for state of E (Editing) in red, J (Jump Editing) in green, P (Long Pause) in blue, and T (Text Production) in cyan.

entering state S_j. The number of states I is finite. In the Guo et al. (2019) study, three writing states were used in modeling students' writing processes for subgroup comparison. One of those states was an editing state which combined qualitatively different types of editing behavior. In the current study, we extended the model to four states so that two important and distinct writing types of editing behavior, local and global, could be differentiated in investigating the effect of lead-in tasks.

In this section, we introduce the necessary notations for understanding the analysis. In a semi-Markov model, the duration time follows a more general distribution than an exponential distribution, such as the Weibull distribution with a scale parameter σ and a shape parameter $\nu \in (0, \infty)$. When $\nu = 1$, the Weibull distribution degenerates to an exponential distribution, and the semi-Markov model becomes a continuous time Markov Chain (CTMC; Jackson 2011). Note the subscripts of states for the above parameters are omitted for simplicity.

In a semi-Markov model, two groups of parameters need to be estimated: one is the transition probability matrix for the embedded Markov chain, and the other is the duration time parameters. Based on these parameters, two hazard rates are defined to characterize a semi-Markov model: the hazard rate of duration time (denoted as $\alpha_{ij}(t)$, which is the likelihood of leaving the current state i for the next state j at time t) and the hazard rate of the process (denoted as $\lambda_{ij}(t)$, which is the instantaneous transition probability from State i to State j at time t). Details can be found in Guo et al. (2019) and Krol and Saint-Pierre (2015).

To compare the essay position (first vs. last) or the mixed-scenario effects on writing processes, we used the Cox proportional regression model (Cox, 1972) in the semi-Markov modeling. The influence of the covariate Z on the hazard rate $\alpha_{ij}(t)$ is denoted by

$$\alpha_{ij1}(t|Z) = \alpha_{ij0}(t) \exp(\beta_{ij} Z), \tag{1}$$

where $\alpha_{ij1}(t)$ and $\alpha_{ij0}(t)$ are the hazard rates of the focal ($Z = 1$) and reference ($Z = 0$) groups at time t, 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. In this study, we use the R-package *SemiMarkov* (Krol and Saint-Pierre, 2015) to fit semi-Markov models and estimate parameters.

4. Results

4.1. Summary Statistics

Table 2 shows the descriptive summary statistics for the four experimental groups, which recapitulate the main findings of (Zhang et al., 2019). The four groups were not measurably different in terms of teacher ratings of writing skill (the third column). The total test scores between the two single-scenario forms (Forms 1 and 2), or between the mixed-scenario forms (Forms 3 and 4), were also not statistically significantly different. However, the total scores on the single-scenario forms were statistically different from those mixed-scenario forms; this difference may be explained by the fact that the lead-in tasks and source materials for those tasks were different, forcing students to switch topical contexts midstream. Essay scores on the two essay-first forms (Form 2 and Form 4) are comparable as well. Essay scores on Forms 2 & 4 were not statistically significantly different from those on the single-scenario & essay-last form (Form 1), but they are much higher than those on the mixed-scenario & essay-last form (Form 3).

Students' total response times and essay writing times on the mixed-scenario & essay-last form (Form 3) were shorter and had less variance compared to the other three forms. Students who took the essay-first forms (Form 2 and Form 4) are more comparable in terms of total response time and essay writing time.

Table 2: Descriptive Summary Statistics for Teacher Rating (TR), Scores, and Response Time by Form

		TR	Total	Task 1	Task 2	Task 3	Essay
			Test & Task Scores				
Form	N	Mean(SD)	Mean(SD)	Mean(SD)	Mean(SD)	Mean(SD)	Mean(SD)
1	257	3.26(.90)	19.19(6.77)	8.17(3.23)	1.18(1.30)	5.36(1.95)	4.50(1.90)
2	261	3.32(.94)	19.20(6.08)	7.84(3.04)	1.07(1.22)	5.49(1.81)	4.82(1.92)
3	271	3.40(.95)	17.41(6.04)	7.25(2.81)	1.14(1.25)	5.51(1.95)	3.52(1.65)
4	260	3.18(.90)	17.02(6.17)	6.55(2.90)	0.88(1.14)	4.95(2.16)	4.66(1.80)
			Response Time (in seconds)				
Form	N		Median(IQR)	Median(IQR)	Median(IQR)	Median(IQR)	Median(IQR)
1	257		2358(1012)	1083(452)	253(210)	241(120)	708(486)
2	261		2415(1051)	995(518)	229(180)	230(106)	944(608)
3	271		2297(770)	1222(426)	253(173)	254(154)	525(247)
4	260		2324(1112)	1044(552)	205(136)	254(94)	895(590)

Note. TR = teacher rating of student writing skill; Essay score = sum of the two rubric scores; IQR = interquartile range.

Table 3: Summary Statistics for Writing States by Form

		Number of Writing States		
	Student N	Total	Mean	SD
Form 1	225	17347	77.10	45.69
Form 2	248	25947	104.63	60.31
Form 3	233	13126	56.33	26.44
Form 4	238	24854	104.43	70.17

Table 3 shows the summary statistics for the total number of writing states on each of the four test forms (note that the sample sizes are slightly smaller than in Table 2 because we removed students whose keystroke logs were not captured correctly). Using Wilcoxon rank sum tests, no measurable differences were detected between Forms 2 and 4 (p-value=.43), but students in the sample taking Form 1, Form 2 and Form 4 had statistically significantly longer logs than students taking Form 3 (p-values are less than .0001).

Table 4: Transition Frequencies (Relative Frequencies) Among Four Writing Process States

		E	J	P	T
Form 1	E	*3112 (100%)*	75 (2%)	653 (21%)	2384 (77%)
	J	30 (5%)	*616 (100%)*	458 (74%)	128 (21%)
	P	577 (10%)	48 (1%)	*5834 (100%)*	5209 (89%)
	T	2518 (33%)	504 (7%)	4538 (60%)	*7560 (100%)*
Form 2	E	*4616 (100%)*	651 (14%)	908 (20%)	3057 (66%)
	J	586 (31%)	*1919 (100%)*	804 (42%)	529 (28%)
	P	684 (8%)	151 (2%)	*8292 (100%)*	7456 (90%)
	T	3366 (31%)	1132 (10%)	6375 (59%)	*10873 (100%)*
Form 3	E	*2484 (100%)*	268 (11%)	377 (15%)	1839 (74%)
	J	198 (27%)	*723 (100%)*	343 (47%)	182 (25%)
	P	339 (8%)	48 (1%)	*4115 (100%)*	3728 (91%)
	T	1963 (35%)	418 (8%)	3190 (57%)	*5571 (100%)*
Form 4	E	*4689 (100%)*	626 (13%)	835 (18%)	3228 (69%)
	J	520 (28%)	*1865 (100%)*	705 (38%)	640 (34%)
	P	723 (10%)	123 (2%)	*7598 (100%)*	6752 (89%)
	T	3468 (33%)	1131 (8%)	5865 (56%)	*10464 (100%)*

Note. On the off-diagonal, the states in the left column are the starting points and the states in the top row are the transition landing points. Statistics on the diagonal are the total (relative) frequencies starting from each of the starting states.

Table 4 shows the transition frequencies and relative frequencies between different writing states for students who took Forms 1, 2, 3, and 4, respectively. Note that the state transitions have direction; the states in the far-left column in the table are the starting points, and the states in the top row are the ending points. For example, in the first panel for Form 1, from E to J, the frequency is 75, which is 2% of the total 3112 transitions from E; from J to E, the frequency is

30, which is 5% of the total 616 transitions from J. The numbers on the diagonal in each panel are the total frequencies of the transitions starting from each of the four states.

We first observed that J was the writing state with the lowest number of occurrences in all four forms, which rendered the smaller relative frequencies in the J column, compared to others. In addition, the relative frequencies for the E-to-J and J-to-E transitions on Form 1 were much smaller than on the other forms, but the relative frequencies for the E-to-T and J-to-P transitions on Form 1 were much larger than on the other forms.

4.2. Model Selection

As discussed in the introduction, we defined the second revision state: the jump editing state (J), in addition to the local editing (E) state, because we think these are two qualitatively different writing behaviors. Therefore, we conducted a 4-state vs. 3-state Semi-Markov model comparison. In addition, because a semi-Markov model is a more complicated one than CTMC, we investigated the scale parameter ν to evaluate whether our data could be modeled by the simpler CTMC model. Note that, because of the similarity of the semi-Markov model-fitting results, only Form 1 results are given below and those for Forms 2 to 4 are presented in Appendix A.

We first a conducted 4-state vs. 3-state Semi-Markov model comparison to evaluate whether we should differentiate J from E globally. We put constraints on the set of distribution parameters, so that J and E are indifferentiable in the 3-state model. Table 5 shows that the 4-state model fits the data significantly better than the 3-state model, and the likelihood-ratio test has a p-value much less than .0001.

Table 5: Comparison of 3-state vs. 4-state Models for Form 1.

NumStates	NumPara	-2loglikelihood	AIC	BIC	χ^2	p-value
3	24	66979.85	67027.85	67109.84	1224.92	$< .0001$
4	32	65754.93	65818.93	65928.25		

Note. The chi-square statistic for testing the two nested models has eight degrees of freedom.

Table 6 shows the estimated parameters of state duration/sojourn time in the 4-state semi-Markov model for Form 1. For the notation used in the subscripts of the parameters, 1=E, 2=J, 3=P, and 4=T. Most of the values estimated for the shape parameter ν of the state duration time are statistically significantly different from 1 after applying Bonferroni correction (except for $\nu_{21}, \nu_{32}, \nu_{42}$, and ν_{43}). Therefore, a semi-Markov model may be preferable to a continuous time Markov model.

Table 6: Semi-Markov Model Fitting for Form 1.

Parameter	Transition	Estimate	SD	LowerCI	UpperCI	H0	p.value
σ_{12}	E $\rightarrow$ J	3.96	0.35	3.27	4.65	1.00	<0.0001
σ_{13}	E $\rightarrow$ P	4.55	0.13	4.29	4.81	1.00	<0.0001
σ_{14}	E $\rightarrow$ T	4.63	0.07	4.48	4.77	1.00	<0.0001
σ_{21}	J $\rightarrow$ E	12.07	2.33	7.51	16.63	1.00	<0.0001
σ_{23}	J $\rightarrow$ P	18.34	1.09	16.20	20.47	1.00	<0.0001
σ_{24}	J $\rightarrow$ T	21.03	3.29	14.58	27.49	1.00	<0.0001
σ_{31}	P $\rightarrow$ E	9.57	0.51	8.58	10.56	1.00	<0.0001
σ_{32}	P $\rightarrow$ J	9.81	1.68	6.52	13.10	1.00	<0.0001
σ_{34}	P $\rightarrow$ T	10.87	0.20	10.49	11.26	1.00	<0.0001
σ_{41}	T $\rightarrow$ E	8.45	0.20	8.06	8.83	1.00	<0.0001
σ_{42}	T $\rightarrow$ J	11.36	0.58	10.22	12.51	1.00	<0.0001
σ_{43}	T $\rightarrow$ P	10.22	0.16	9.90	10.53	1.00	<0.0001
ν_{12}	E $\rightarrow$ J	1.38	0.12	1.14	1.61	1.00	0.0019
ν_{13}	E $\rightarrow$ P	1.44	0.04	1.36	1.52	1.00	<0.0001
ν_{14}	E $\rightarrow$ T	1.37	0.02	1.34	1.40	1.00	<0.0001
ν_{21}	J $\rightarrow$ E	1.00	0.14	0.73	1.28	1.00	1.0000
ν_{23}	J $\rightarrow$ P	0.83	0.02	0.79	0.88	1.00	<0.0001
ν_{24}	J $\rightarrow$ T	0.60	0.03	0.53	0.67	1.00	<0.0001
ν_{31}	P $\rightarrow$ E	0.84	0.02	0.79	0.88	1.00	<0.0001
ν_{32}	P $\rightarrow$ J	0.90	0.09	0.73	1.07	1.00	0.2435
ν_{34}	P $\rightarrow$ T	0.81	0.01	0.80	0.83	1.00	<0.0001
ν_{41}	T $\rightarrow$ E	0.92	0.01	0.89	0.95	1.00	<0.0001
ν_{42}	T $\rightarrow$ J	0.95	0.03	0.88	1.02	1.00	0.1573
ν_{43}	T $\rightarrow$ P	1.00	0.01	0.97	1.02	1.00	0.7642

Figure 2 shows the estimated density distributions of duration time for the transitions in Table 6, compared to empirical data. This figure shows that the fit of estimated distributions is reasonable, except for the J state because of rare occurrences of J. The sojourn times have a very skewed distribution with a long tail on the right; the median durations for E, J, P, and T were about 3 - 4, 8 - 12, 6 - 7, and 6 - 8 seconds, respectively.

Results from all four groups' writing processes show that semi-Markov processes were acceptable and better than CTMC models in terms of fit to the data, and that the E and J states bear different characteristics from one another. (Tables are available upon request.)

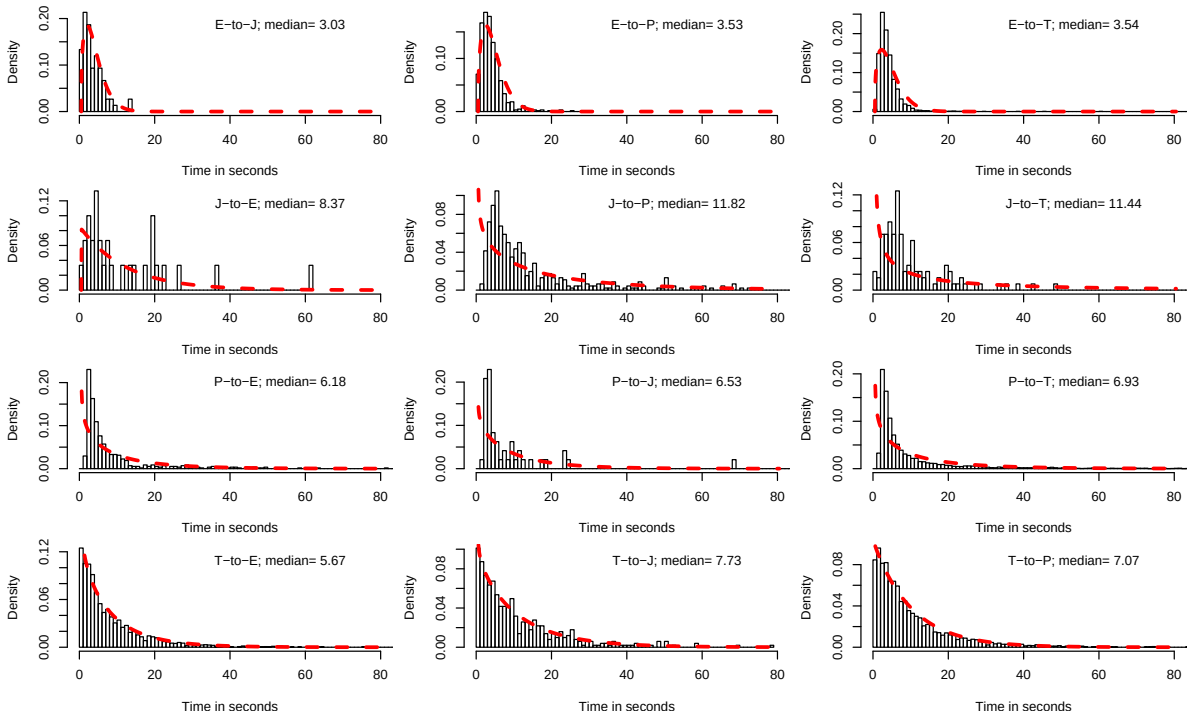

Figure 2: Estimated density distributions of duration time in the Semi-Markov model compared to empirical data for Form 1. In each panel, the bars were produced by frequency histogram, and the red dashed line by the semi-Markov model.

4.3. Essay Position Effect

Because of the random assignment of test forms within a class, we would expect that the students in Groups 2 and 4 (students who took essay-first forms) would have similar levels of writing skill and show similar writing processes. Indeed, their teacher ratings were not measurably different (refer to Tables 2 & 3). In addition, results from semi-Markov modeling (see Appendix) were comparable across the groups. As a consequence, we combined the two groups (denoted as Group 24) for analysis hereafter. In this section, we compare students who took the original single-scenario & essay-last form (Form 1), denoted as Group 1, to Group 24 in their writing processes to evaluate the essay-position effect. In a subsequent section, we compare the two essay-last groups, Group 1 and Group 3 (which took Form 3), to evaluate the effect of mixed-scenario on writing processes.

Figure 3 shows, clockwise from top left, the comparison of Group 1 and Group 24 on the in-word typing speed, essay scores, total essay response time, and total number of words in the submitted essay. The figure clearly shows that Group 1 and Group 24 are quite comparable on their keyboarding skills, and that the essay position did not impact their essay scores. However, as indicated from Figure 3 and two-sample Wilcoxon rank tests, the mean total essay writing time and the mean total number of words (essay lengths) were statistically significantly smaller (shorter) for Group 1 than for Group 24.

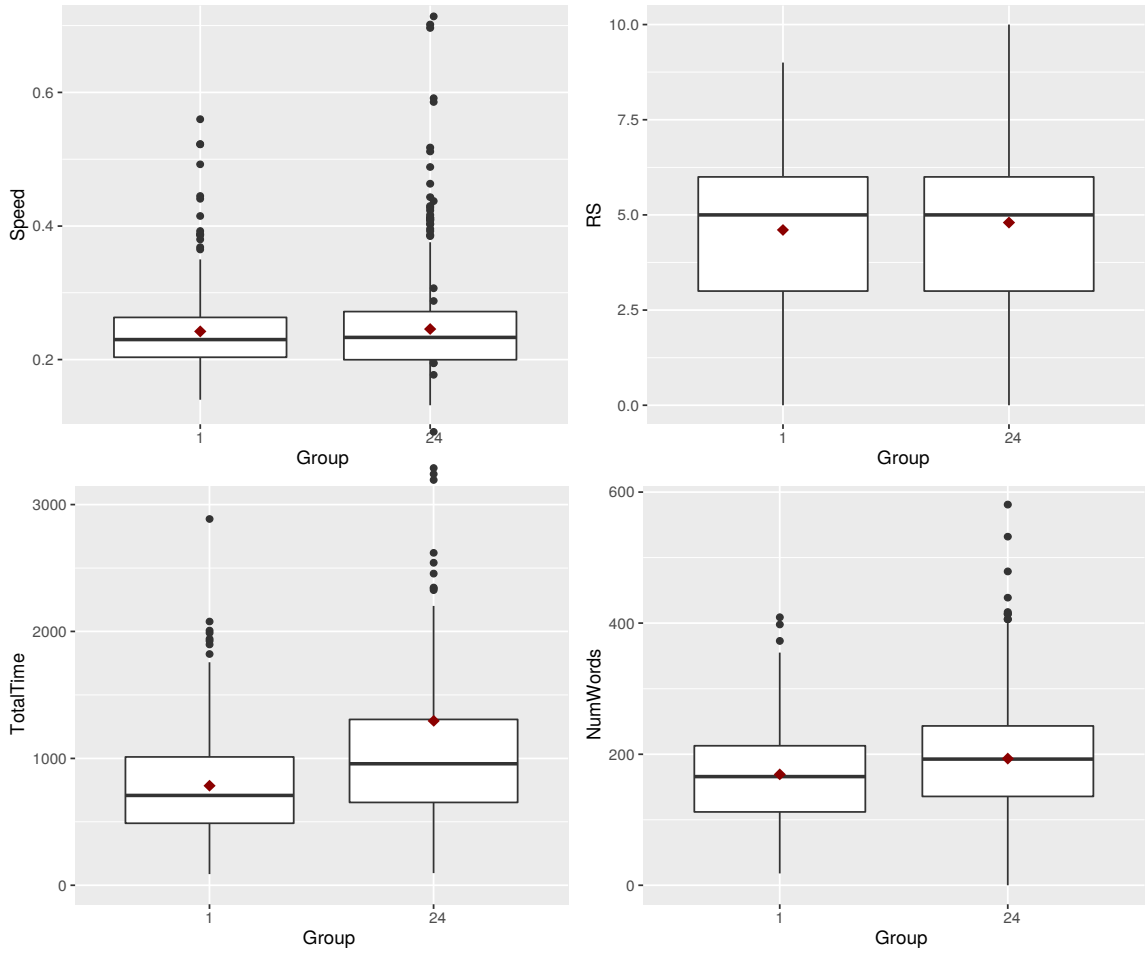

Figure 3: Comparison between Group 1 (left-hand box plot in each panel) and Group 24 (right-hand box plot in each panel). From top-left panel and clockwise, the box plots show in-word typing speed, essay score, total essay time, and total number of words in essay.

In the following analyses, we investigate the differences in writing processes between Group 1 and Group 24 as an attempt to understand better how the two groups undertook the composition task.

Table 7 shows the parameter estimates from the semi-Markov model. From Table 7, we observe that, again, the semi-Markov model fit the data better in view of the scale parameters νs. This finding is largely consistent with the results discussed above on Table 6. Hence, we do not discuss the results pertaining to the νs and σs further.

Table 7: Semi-Markov Model Fitting for Group 1 and Group 24.

Parameter	Transition	Estimation	SD	LowerCI	UpperCI	H0	p.value
σ_{12}	E $\rightarrow$ J	2.25	0.06	2.13	2.37	1.00	<0.0001
σ_{13}	E $\rightarrow$ P	4.46	0.08	4.31	4.62	1.00	<0.0001
σ_{14}	E $\rightarrow$ T	4.26	0.04	4.19	4.33	1.00	<0.0001
σ_{21}	J $\rightarrow$ E	3.87	0.18	3.52	4.23	1.00	<0.0001
σ_{23}	J $\rightarrow$ P	24.13	0.96	22.24	26.02	1.00	<0.0001
σ_{24}	J $\rightarrow$ T	8.24	0.48	7.30	9.19	1.00	<0.0001
σ_{31}	P $\rightarrow$ E	10.25	0.37	9.53	10.97	1.00	<0.0001
σ_{32}	P $\rightarrow$ J	12.42	1.33	9.82	15.02	1.00	<0.0001
σ_{34}	P $\rightarrow$ T	10.69	0.13	10.43	10.95	1.00	<0.0001
σ_{41}	T $\rightarrow$ E	7.83	0.11	7.62	8.04	1.00	<0.0001
σ_{42}	T $\rightarrow$ J	7.05	0.21	6.64	7.46	1.00	<0.0001
σ_{43}	T $\rightarrow$ P	9.36	0.09	9.18	9.53	1.00	<0.0001
ν_{12}	E $\rightarrow$ J	1.07	0.02	1.03	1.11	1.00	0.0008
ν_{13}	E $\rightarrow$ P	1.45	0.02	1.41	1.49	1.00	<0.0001
ν_{14}	E $\rightarrow$ T	1.54	0.01	1.52	1.56	1.00	<0.0001
ν_{21}	J $\rightarrow$ E	0.68	0.01	0.66	0.71	1.00	<0.0001
ν_{23}	J $\rightarrow$ P	0.68	0.01	0.66	0.70	1.00	<0.0001
ν_{24}	J $\rightarrow$ T	0.53	0.01	0.51	0.55	1.00	<0.0001
ν_{31}	P $\rightarrow$ E	0.79	0.01	0.77	0.81	1.00	<0.0001
ν_{32}	P$\rightarrow$ J	0.63	0.02	0.59	0.67	1.00	<0.0001
ν_{34}	P $\rightarrow$ T	0.71	0.00	0.70	0.72	1.00	<0.0001
ν_{41}	T $\rightarrow$ E	0.94	0.01	0.92	0.95	1.00	<0.0001
ν_{42}	T $\rightarrow$ J	0.76	0.01	0.74	0.79	1.00	<0.0001
ν_{43}	T $\rightarrow$ P	0.99	0.01	0.97	1.00	1.00	0.0087
β_{12}	E $\rightarrow$ J	**-0.54**	0.12	-0.77	-0.30	0.00	<0.0001
β_{13}	E $\rightarrow$ P	-0.03	0.05	-0.12	0.06	0.00	0.5323
β_{14}	E $\rightarrow$ T	**-0.21**	0.02	-0.26	-0.17	0.00	<0.0001
β_{21}	J $\rightarrow$ E	**-0.73**	0.19	-1.11	-0.36	0.00	0.0001
β_{23}	J $\rightarrow$ P	**0.27**	0.05	0.16	0.38	0.00	<0.0001
β_{24}	J $\rightarrow$ T	**-0.45**	0.09	-0.63	-0.26	0.00	<0.0001
β_{31}	P $\rightarrow$ E	0.09	0.05	-0.01	0.18	0.00	0.0793
β_{32}	P $\rightarrow$ J	0.25	0.16	-0.06	0.57	0.00	0.1131
β_{34}	P $\rightarrow$ T	**0.05**	0.02	0.02	0.08	0.00	0.0026
β_{41}	T $\rightarrow$ E	**-0.07**	0.02	-0.12	-0.03	0.00	0.0022
β_{42}	T $\rightarrow$ J	**-0.35**	0.05	-0.45	-0.24	0.00	<0.0001
β_{43}	T $\rightarrow$ P	**-0.08**	0.02	-0.11	-0.04	0.00	<0.0001

The results that are of most interest relate to the estimates of the hazard ratio β_{ij} parameter, which are indicative of group differences. For the group comparison, Group 1 is treated as the focal group, and the parameters β_{ij} show that, on average, the duration time in E and in T (and possibly in J) of Group 1 was longer than that of Group 24 ($\beta_{1j} < 0$ and $\beta_{4j} < 0$). In addition, in Table 7, (and later Table 8 as well), some βs (boldfaced values) are statistically different from zero after Bonferroni correction. More specifically, compared to Group 24, Group 1 spent longer

time in E before transiting to J or T ($\beta_{12} = -.54$ and $\beta_{14} = -.21$), spent longer time in J before transiting to E or T ($\beta_{21} = -.73$) and ($\beta_{24} = -.45$), and spent longer time in T before transiting to J (β_{42}=-.35). On the other hand, β_{23} (.27) is significantly larger than 0, indicating that Group 1 spent less time in J before transiting to P.

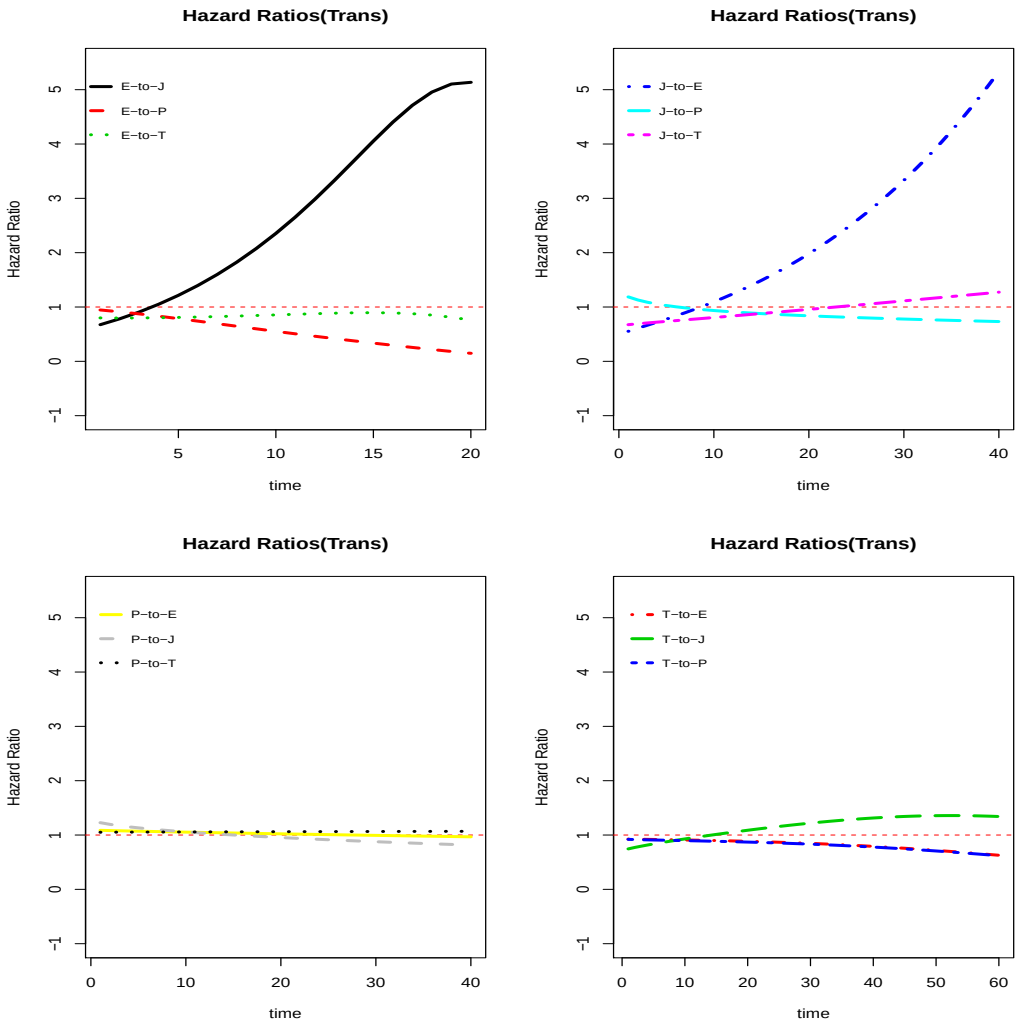

Figure 4: Hazard Ratios (transition) for Group 1 Compared with Group 24. From top left going clockwise, transition, from E, from J, from T, and from P.

Figure 4 shows that, compared to Group 24, Group 1 students were more (or less) likely to transition to J (or P) from E after about three seconds. From J, they were more likely to transition to E after about ten seconds. From T, they were more likely to transition to J after about ten seconds and less likely to transition to E & P after about ten seconds.

4.4. Effect of Single Scenario vs. Mixed Scenarios

In this section, we compared the writing-process patterns between the single-scenario & essay last group (Group 1) with the mixed-scenario & essay-last group (Group 3). Figure 5 shows, clockwise from top left, the comparison of Group 1 and Group 3 on in-word typing speed, essay scores, total essay response time, and total number of words in the essay.

Figure 5 shows that Group 1 and Group 3 are quite comparable in typing speed. However, similar to findings by Zhang, van Rijn, et al. (2019), the mixed-scenario condition experienced by Group 3 resulted in statistically significantly lower mean essay scores. In addition, on average, the total essay writing time and the total number of words were statistically significantly larger for Group 1 than for Group 3 (Wilcoxon rank tests have p-values less than .0001). Zhang, van Rijn, et al. (2019) suggested that these differences might be due to lower motivation caused by the introduction of a second scenario and the second set of source materials for the essay that concluded the test.

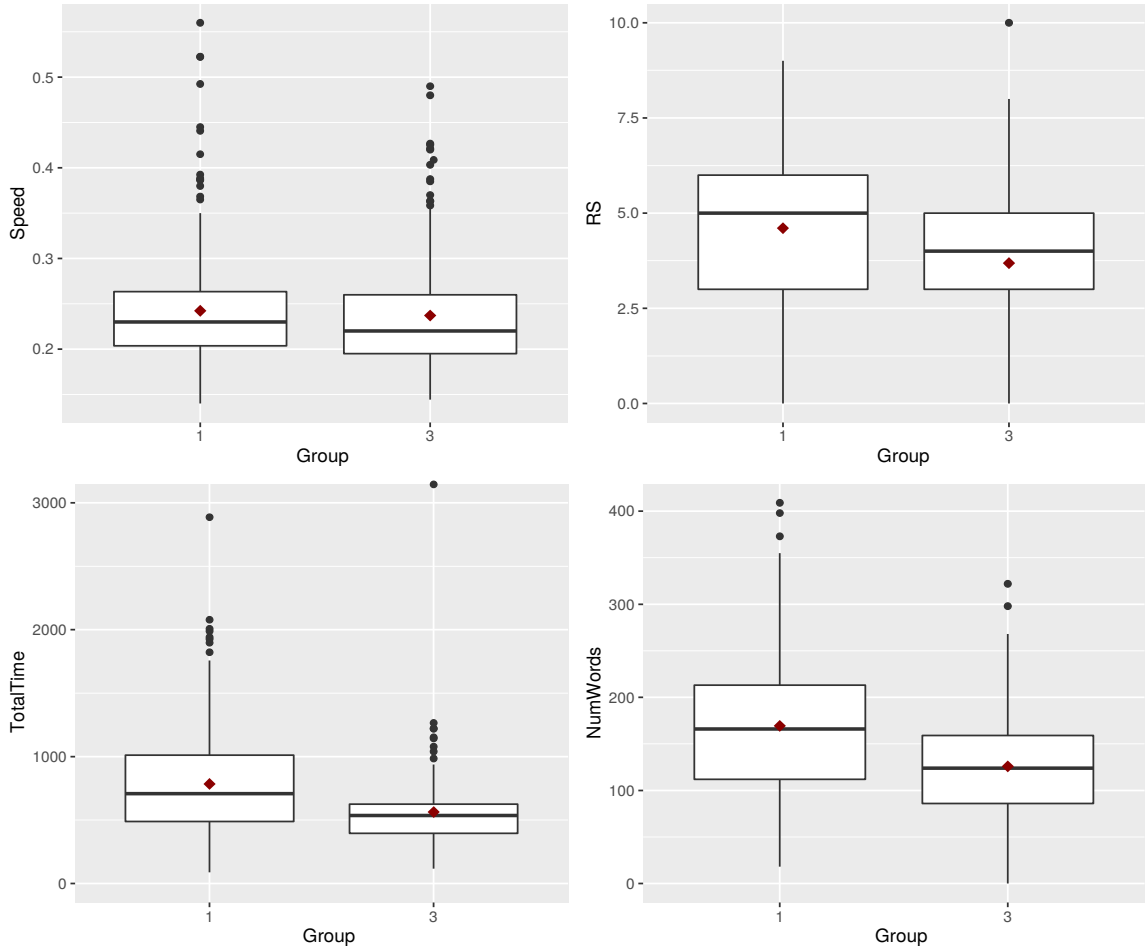

Figure 5: Clockwise, from top-left, in-word typing speed, essay score, total essay time and total number of words in the essay, respectively, for Group 1 vs. Group 3.

Table 8: Semi-Markov Model Fitting for Group 1 and Group 3.

Label	Transition	Estimation	SD	LowerCI	UpperCI	H0	p.value
σ_{12}	E $\rightarrow$ J	2.42	0.11	2.21	2.64	1.00	<0.0001
σ_{13}	E $\rightarrow$ P	4.00	0.14	3.73	4.27	1.00	<0.0001
σ_{14}	E $\rightarrow$ T	4.15	0.07	4.01	4.28	1.00	<0.0001
σ_{21}	J $\rightarrow$ E	3.25	0.31	2.65	3.85	1.00	<0.0001
σ_{23}	J $\rightarrow$ P	21.55	1.71	18.20	24.90	1.00	<0.0001
σ_{24}	J $\rightarrow$ T	10.18	1.23	7.77	12.59	1.00	<0.0001
σ_{31}	P $\rightarrow$ E	8.60	0.55	7.53	9.68	1.00	<0.0001
σ_{32}	P $\rightarrow$ J	10.40	1.77	6.92	13.88	1.00	<0.0001
σ_{34}	P $\rightarrow$ T	10.34	0.22	9.91	10.77	1.00	<0.0001
σ_{41}	T $\rightarrow$ E	8.77	0.22	8.33	9.21	1.00	<0.0001
σ_{42}	T $\rightarrow$ J	9.91	0.54	8.85	10.97	1.00	<0.0001
σ_{43}	T $\rightarrow$ P	10.30	0.19	9.92	10.67	1.00	<0.0001
ν_{12}	E $\rightarrow$ J	1.44	0.06	1.32	1.55	1.00	<0.0001
ν_{13}	E $\rightarrow$ P	1.51	0.03	1.44	1.57	1.00	<0.0001
ν_{14}	E $\rightarrow$ T	1.46	0.01	1.43	1.48	1.00	<0.0001
ν_{21}	J $\rightarrow$ E	0.80	0.03	0.74	0.87	1.00	<0.0001
ν_{23}	J $\rightarrow$ P	0.72	0.02	0.69	0.75	1.00	<0.0001
ν_{24}	J $\rightarrow$ T	0.64	0.02	0.59	0.68	1.00	<0.0001
ν_{31}	P $\rightarrow$ E	0.87	0.02	0.83	0.90	1.00	<0.0001
ν_{32}	P $\rightarrow$ J	0.88	0.06	0.76	1.00	1.00	0.0469
ν_{34}	P $\rightarrow$ T	0.80	0.01	0.79	0.81	1.00	<0.0001
ν_{41}	T $\rightarrow$ E	0.94	0.01	0.92	0.96	1.00	<0.0001
ν_{42}	T $\rightarrow$ J	0.97	0.03	0.92	1.03	1.00	0.3197
ν_{43}	T $\rightarrow$ P	0.99	0.01	0.98	1.01	1.00	0.4751
β_{12}	E $\rightarrow$ J	**-0.73**	0.13	-0.99	-0.47	0.00	<0.0001
β_{13}	E $\rightarrow$ P	**-0.22**	0.07	-0.35	-0.09	0.00	0.0009
β_{14}	E $\rightarrow$ T	**-0.20**	0.03	-0.26	-0.14	0.00	<0.0001
β_{21}	J $\rightarrow$ E	**-1.02**	0.20	-1.41	-0.62	0.00	<0.0001
β_{23}	J $\rightarrow$ P	0.18	0.07	0.04	0.32	0.00	0.0128
β_{24}	J $\rightarrow$ T	**-0.50**	0.12	-0.73	-0.26	0.00	<0.0001
β_{31}	P $\rightarrow$ E	-0.12	0.07	-0.25	0.02	0.00	0.0937
β_{32}	P $\rightarrow$ J	0.06	0.20	-0.34	0.46	0.00	0.7642
β_{34}	P $\rightarrow$ T	-0.03	0.02	-0.07	0.01	0.00	0.1345
β_{41}	T $\rightarrow$ E	0.03	0.03	-0.04	0.09	0.00	0.3994
β_{42}	T $\rightarrow$ J	-0.14	0.07	-0.28	-0.00	0.00	0.0439
β_{43}	T $\rightarrow$ P	0.01	0.02	-0.04	0.05	0.00	0.7184

From Table 8, we observed that, on average, the duration time in E (and possibly in J) of Group 1 was longer than that of Group 3 ($\beta_{1\cdot} < 0$). Some βs (boldfaced values) are significantly different from zero (after Bonferroni correction). In particular, $\beta_{12} = -.73$, $\beta_{21} = -1.02$, and $\beta_{24} = -.50$ are significantly smaller statistically than 0, indicating that Group 1 spent a longer time in E before the transition to J and in J before the transition to T. Duration times on T and P mostly did not show statistically significant differences between Group 1 and Group 3.

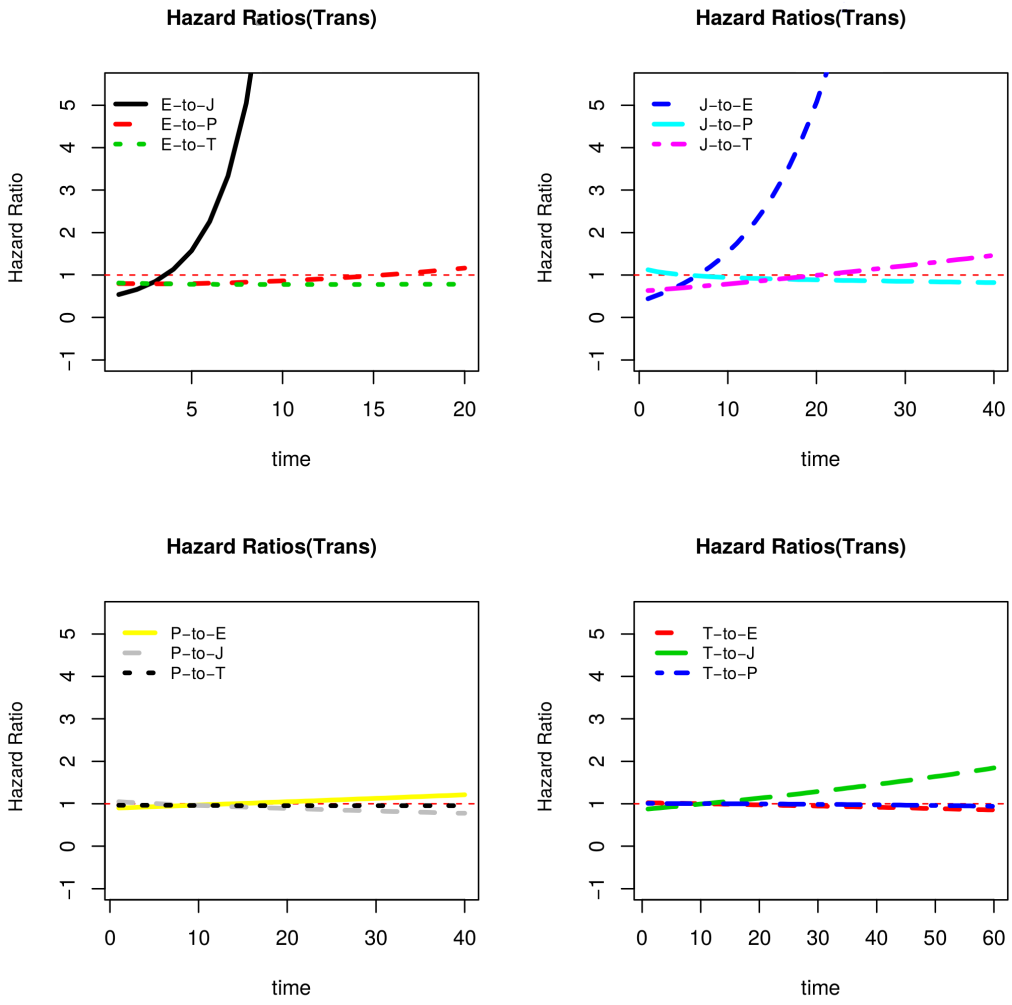

Figure 6: Hazard ratios (transition) for Group 1 and Group 3. From top left going clockwise, transition from E, from J, from T, and from P.

Figure 6 shows the hazard ratios of transition. Compared to Group 3, Group 1 students were more likely to transition to J from E after about three seconds; they were more likely to transition to E from J after about ten seconds; and they were more likely to transition to J from T after about 20 seconds. Of note, is that all of these differences between the two groups relate to transitions to or from J.

5. Discussion

The current study complemented a prior investigation that evaluated the impact of a scenario-based assessment design on test scores and on the psychometric characteristics of those scores. Using data gathered through that prior investigation, we evaluated the impact of that scenario-based assessment structure on students' writing processes. In particular, we compared four assessment design conditions in terms of the writing states that students evidenced in composing their essays and the time they spent in those states. Keystroke logs were used to classify the

actions occurring over the course of composition into: a planning or long pause state (P); a state that combines translation and transcription processes which we denote as text production (T); and two revision states designated as local editing (E) and jump editing (J). Semi-Markov models were then fit to the data so that the four writing states and their duration times could be compared among the different assessment designs.

On average, as expected, students' jump editing states were rare in all four forms compared to editing, long pause, and text production states. However, jump editing states were much longer on average and they are more spread out than the local editing states, which may indicate that jump edits are likely to be linked to global revision processes. Compared to those who took the other three forms, students who took the original single-scenario & essay-last design (Group 1) had relatively fewer (but longer) jump editing states; after a jump editing state, they transitioned more frequently to a long pause state. After a local editing state, students in Group 1 transitioned more frequently into a text production state, and their local editing states were relatively longer as well. In addition, students in Group 1 had a higher tendency to transition between long-duration local editing and jump editing states than students in the other three groups.

As in the prior study by Zhang et al. (2019), we found that the position of the essay itself did not seem to have a significant impact on essay scores, and that the placement of lead-in tasks prior to the essay enabled students to produce essays similar in quality in less time using fewer words. But we also found that students who took the original single-scenario & essay-last design (Group 1) had longer duration in the editing and text production states and shorter duration in the long pause state transiting to text production than students who wrote the essay first without the benefit of the lead-in tasks (Group 24). Students in Group 1 had a higher tendency to make transitions related to longer jump editing states (around ten seconds or so), and a lower tendency to make transitions to longer long pause states from other states than those in Group 24.

Compared to those who took the mixed-scenario & essay-last form (Group 3), students who took the original single-scenario & essay-last form (Group 1) wrote better essays. In their writing processes, students in Group 1 had duration times in long pause states and text production states similar to Group 3, but they seemed to also have a longer duration in local editing and jump editing states compared to Group 3. In addition, students in Group 1 had a higher tendency to make jump-related transitions (E $\rightarrow$ J; J $\rightarrow$ E, and T $\rightarrow$ J) after long state duration, compared to students in Group 3.

Overall, the original single-scenario & essay-last form with the theoretically motivated assessment design may have reduced students' working memory load while they planned and reviewed their essays. Decreased working-memory load during writing is likely to reduce the frequency and duration of long pauses while increasing burst length (Deane et al., 2018) and increasing the attention available for monitoring and revising the text produced. The result would be more efficient and fluent text production, freeing time and energy to scan the text to monitor and correct problems. This explanation accounts for Group 1 students' increased time in the text production state, the longer duration of their jump editing state, and their increased likelihood of switching back and forth between jumps and edits when those states lasted for more than five to ten seconds. The contrasting pattern observed for Group 24, the essay-first group, would on this account be due primarily to the added cognitive load of having to read and process the source information without the benefit of the preparation afforded by the lead-in questions. This added cognitive load would result in students spending more time on task in order to write an essay they considered good enough to submit.

For students in Group 3 who took the mixed-scenario & essay-last form, the situation is similar. Students in Group 3 produced much worse essays for the same levels of text production compared to students in Group 1 and were less likely to transition between jump and edit states after a long duration. Students in Group 3 still had to read the relevant source texts to write an essay, but also had to suppress information about an unrelated topic that was the focus of the preceding lead-in questions. This additional working-memory load, possibly combined with reduced motivation, presumably accounts for the lower essay scores and the lower prevalence of editing. Note that since Group 3 essays were shorter than Group 1 essays, students in Group 3 were necessarily using their text production time less fluently and productively than students in Group 1.

In our data, jump-editing behaviors were relatively rare, and therefore relatively difficult to model. However, applying semi-Markov modeling enabled us to detect significant differences between the study groups on these writing states, and it highlighted a behavior (i.e., longer duration associated with a pattern of alternating between jumps and edits), which may be an important indicator of effort being put into monitoring and correcting text. This result suggests that semi-Markov modeling of writing processes may help researchers better understand students' writing behaviors, helping to map observable events recorded by the keystroke logs onto states interpretable by cognitive theories of writing.

In our analysis, we used relatively simple heuristics to define our writing states under the guidance of cognitive writing theory. Further research that combines think-aloud protocols and other sources of evidence with keystroke logging might enable us to build even more effective classification procedures, resulting in more accurate modeling of how students switch between writing states. In addition, advanced methods such as machine learning (Uto et al., 2020) can be investigated for finding latent writing states so as to enrich existing cognitive writing theory and analysis methodologies.

It should be noted that both our data and our methods impose important limitations on the conclusions that we can draw. As is the case for all educational and psychological research, results may not necessarily generalize to a broader population or to different conditions so replication with other conditions and participant samples is typically recommended. In the current case, because we observed a fairly large impact of the mixed-scenario design, as well as a more subtle effect from lead-in tasks, our results suggest directions worth exploring, such as examining the relation between the common practice on standardized tests of topic switching and motivation, and how the impact of lead-in tasks might be mediated by the writer's degree of use of source materials. From a technical perspective, as noted by Guo et al. (2019), our current modeling methods do not address the fact that writing behaviors, such as editing, may change in probability and contribution to essay quality over the course of a writing session (Breetvelt et al., 1994). One might, in future studies, try more complex and dynamic state models, but this effort will involve collecting and annotating a larger sample of keystroke logs so that one can estimate how the probabilities of entering into specific writing states change over time.

6. Acknowledgments

We would like to thank Matt Johnson, Rebecca Zwick, Peter van Rijn, and Paul Jewsbury for their helpful comments on an earlier version of the paper. Any opinions expressed in this publication are those of the authors and not necessarily of Educational Testing Service.

REFERENCES

BAAIJEN, V. M., GALBRAITH, D., AND DE GLOPPER, K. 2012. Keystroke analysis: Reflections on procedures and measures. *Written Communication 29,* 3, 246–277.

BENNETT, R. E., DEANE, P., AND VAN RIJN, P. W. 2016. From cognitive domain theory to assessment practice. *Educational Psychologist 51*, 82–107.

BREETVELT, I., VAN DEN BERGH, H., AND RIJLAARSDAM, G. 1994. Relations between writing processes and text quality: When and how? *Cognition and Instruction 12,* 2, 103–123.

COMMON CORE STATE STANDARDS INITIATIVE. 2010. Common core state standards for English language arts and literacy in history/social studies, science, science, and technical subjects.

COX, D. R. 1972. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological) 34,* 2, 187–220.

DEANE, P., FOWLES, M., BALDWIN, D., AND PERSKY, H. 2011. The CBAL summative writing assessment: A draft eighth-grade design. *Research Memorandum No. RM-11-01). Princeton, NJ: Educational Testing Service*.

DEANE, P., ODENDAHL, N., QUINLAN, T., FOLWES, M., WELSH, C., AND BIVENS-TATUM, J. 2018. Cognitive models of writing: Writing proficiency as a complex integrated skill. *ETS Research Report RR-08-55*, 1–120.

DEANE, P., SONG, Y., VAN RIJN, P., O'REILLY, T., FOWLES, M., BENNETT, R., SABATINI, J., AND ZHANG, M. 2019. The case for scenario-based assessment of written argumentation. *Reading and Writing 32,* 6 (Jun), 1575–1606.

DEANE, P. AND ZHANG, M. 2015. Exploring the feasibility of using writing process features to assess text production skills. *ETS Research Report Series 2015,* 2, 1–16.

GUO, H., DEANE, P. D., VAN RIJN, P. W., ZHANG, M., AND BENNETT, R. E. 2018. Modeling basic writing processes from keystroke logs. *Journal of Educational Measurement 55,* 2, 194–216.

GUO, H., ZHANG, M., DEANE, P., AND BENNETT, R. E. 2019. Writing process differences in subgroups reflected in keystroke logs. *Journal of Educational and Behavioral Statistics 44,* 5, 571–596.

HAYES, J. R. 2012. Modeling and remodeling writing. *Written Communication 29,* 3, 369–388.

JACKSON, C. 2011. Multi-state models for panel data: The msm package for R. *Journal of Statistical Software, Articles 38,* 8, 1–28.

KELLOGG, R. T. 2001. Competition for working memory among writing processes. *The American Journal of Psychology 114,* 2, 175–191.

KROL, A. AND SAINT-PIERRE, P. 2015. SemiMarkov: An R package for parametric estimation in multi-state semiMarkov models. *Journal of Statistical Software, Articles 66,* 6, 1–16.

LEIJTEN, M. AND VAN WAES, L. 2013. Keystroke logging in writing research: Using Inputlog to analyze and visualize writing processes. *Written Communication 30,* 3, 358–392.

UTO, M., MIYAZAWA, Y., KATO, Y. ANDNAKAJIMA, K., AND KUWATA, H. 2020. Time- and learner-dependent hidden Markov model for writing process analysis using keystroke log data. *Int J Artif Intell Educ*, 1–28.

ZHANG, M., BENNETT, R. E., DEANE, P., AND VAN RIJN, P. W. 2019. Are there gender differences in how students write their essays? An analysis of writing processes. *Educational Measurement: Issues and Practice 38,* 2, 14–26.

ZHANG, M., HAO, J., LI, C., AND DEANE, P. 2018. Defining personalized writing burst measures of translation using keystroke logs. In *Proceedings of the 11th International Conference on Educational Data Mining*, K. E. Boyer and M. Yudelson, Eds. Buffalo, NY, 549–552.

ZHANG, M., VAN RIJN, P. W., DEANE, P., AND BENNETT, R. E. 2019. Scenario-based assessments in writing: An experimental study. *Educational Assessment 24,* 2, 73–90.

ZHANG, M., ZHU, M., DEANE, P., AND GUO, H. 2019. Identifying and comparing writing process patterns using keystroke logs. In *Quantitative Psychology*, M. Wiberg, S. Culpepper, R. Janssen, J. González, and D. Molenaar, Eds. Springer International Publishing, Cham, 367–381.

ZHU, M., ZHANG, M., AND DEANE, P. 2019. Analysis of keystroke sequences in writing logs. *ETS Research Report Series*.

7. APPENDIX: MODEL FITTING FOR INDIVIDUAL FORMS

Table 9: Semi-Markov Model Fitting for Form 2. There were 254 students in the sample with essay keystroke data.

Label	Transition	Estimation	SD	LowerCI	UpperCI	H0	p.value
σ_{12}	E -> J	2.15	0.09	1.98	2.33	1.00	<0.0001
σ_{13}	E -> P	4.49	0.11	4.26	4.71	1.00	<0.0001
σ_{14}	E -> T	4.34	0.05	4.25	4.43	1.00	<0.0001
σ_{21}	J -> E	3.77	0.23	3.32	4.22	1.00	<0.0001
σ_{23}	J -> P	22.07	1.30	19.52	24.61	1.00	<0.0001
σ_{24}	J -> T	9.27	0.74	7.82	10.71	1.00	<0.0001
σ_{31}	P -> E	10.00	0.54	8.94	11.05	1.00	<0.0001
σ_{32}	P -> J	12.80	2.02	8.84	16.76	1.00	<0.0001
σ_{34}	P -> T	10.54	0.19	10.15	10.92	1.00	<0.0001
σ_{41}	T -> E	7.92	0.16	7.62	8.23	1.00	<0.0001
σ_{42}	T -> J	7.86	0.31	7.24	8.47	1.00	<0.0001
σ_{43}	T -> P	9.63	0.13	9.37	9.88	1.00	<0.0001
ν_{12}	E -> J	1.01	0.03	0.95	1.07	1.00	0.6985
ν_{13}	E -> P	1.38	0.03	1.32	1.44	1.00	<0.0001
ν_{14}	E -> T	1.75	0.02	1.71	1.80	1.00	<0.0001
ν_{21}	J -> E	0.72	0.02	0.69	0.76	1.00	<0.0001
ν_{23}	J -> P	0.64	0.01	0.61	0.66	1.00	<0.0001
ν_{24}	J -> T	0.58	0.02	0.55	0.61	1.00	<0.0001
ν_{31}	P -> E	0.76	0.02	0.73	0.79	1.00	<0.0001
ν_{32}	P -> J	0.55	0.03	0.50	0.60	1.00	<0.0001
ν_{34}	P -> T	0.66	0.00	0.66	0.67	1.00	<0.0001
ν_{41}	T -> E	0.94	0.01	0.91	0.96	1.00	<0.0001
ν_{42}	T -> J	0.80	0.02	0.76	0.84	1.00	<0.0001
ν_{43}	T -> P	0.98	0.01	0.97	1.00	1.00	0.1245

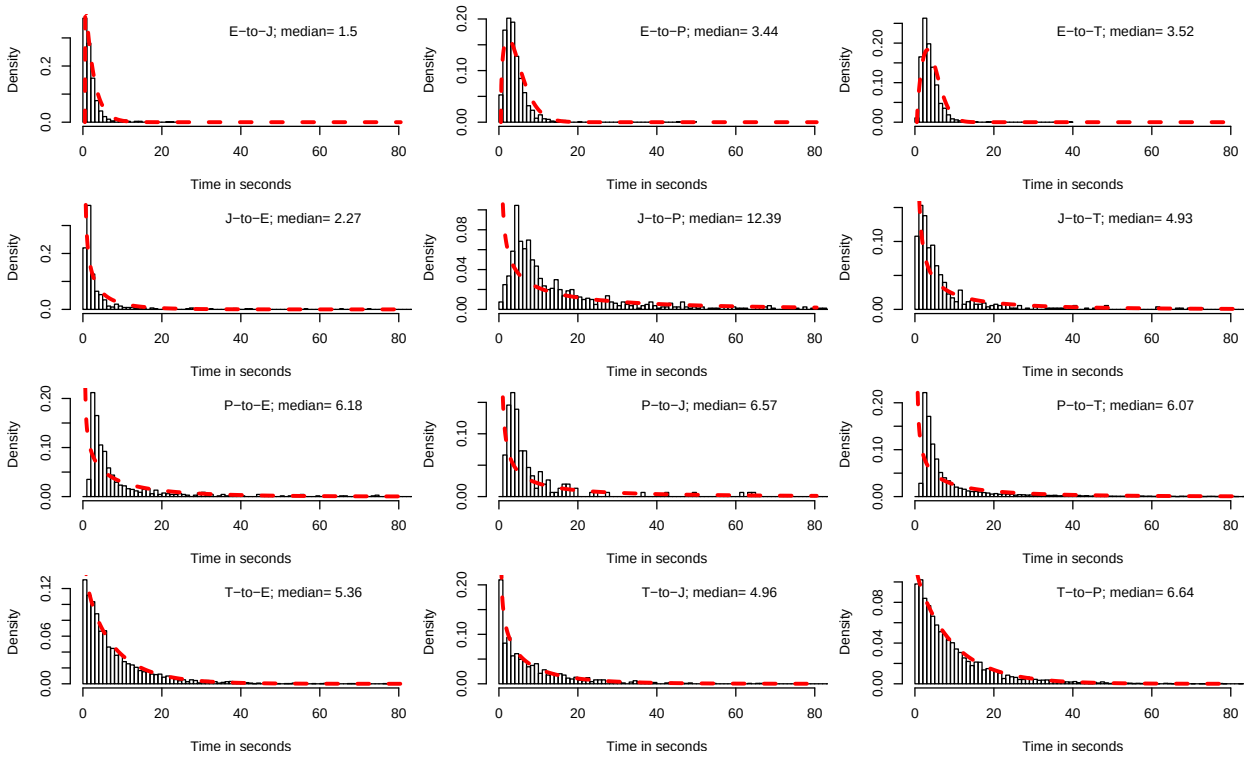

Figure 7: Estimated density distributions of sojourn time in the semi-Markov model compared to empirical data for Form 2. In each panel, the bars were produced by histogram, and the red dashed line by the semi-Markov model.

Table 10: Semi-Markov Model Fitting for Form 3. There were 236 students in the sample with essay keystroke data.

Label	Transition	Estimation	SD	LowerCI	UpperCI	H0	p.value
σ_{12}	E -> J	2.44	0.11	2.22	2.65	1.00	<0.0001
σ_{13}	E -> P	4.11	0.13	3.85	4.37	1.00	<0.0001
σ_{14}	E -> T	4.30	0.06	4.17	4.43	1.00	<0.0001
σ_{21}	J -> E	3.22	0.31	2.61	3.83	1.00	<0.0001
σ_{23}	J -> P	18.84	1.68	15.54	22.14	1.00	<0.0001
σ_{24}	J -> T	10.68	1.28	8.17	13.18	1.00	<0.0001
σ_{31}	P -> E	9.01	0.55	7.93	10.09	1.00	<0.0001
σ_{32}	P -> J	10.30	1.85	6.67	13.93	1.00	<0.0001
σ_{34}	P -> T	10.19	0.23	9.75	10.63	1.00	<0.0001
σ_{41}	T -> E	8.87	0.22	8.44	9.31	1.00	<0.0001
σ_{42}	T -> J	10.03	0.54	8.97	11.08	1.00	<0.0001
σ_{43}	T -> P	10.28	0.20	9.89	10.66	1.00	<0.0001
ν_{12}	E -> J	1.45	0.06	1.32	1.58	1.00	<0.0001
ν_{13}	E -> P	1.67	0.06	1.55	1.79	1.00	<0.0001
ν_{14}	E -> T	1.64	0.02	1.60	1.69	1.00	<0.0001
ν_{21}	J -> E	0.79	0.04	0.72	0.86	1.00	<0.0001
ν_{23}	J -> P	0.64	0.02	0.60	0.68	1.00	<0.0001
ν_{24}	J -> T	0.67	0.03	0.60	0.73	1.00	<0.0001
ν_{31}	P -> E	0.95	0.03	0.88	1.02	1.00	0.1277
ν_{32}	P -> J	0.86	0.09	0.69	1.03	1.00	0.1016
ν_{34}	P -> T	0.79	0.01	0.77	0.80	1.00	<0.0001
ν_{41}	T -> E	0.97	0.02	0.94	1.00	1.00	0.0793
ν_{42}	T -> J	1.00	0.04	0.93	1.08	1.00	1.0000
ν_{43}	T -> P	0.99	0.01	0.96	1.02	1.00	0.4348

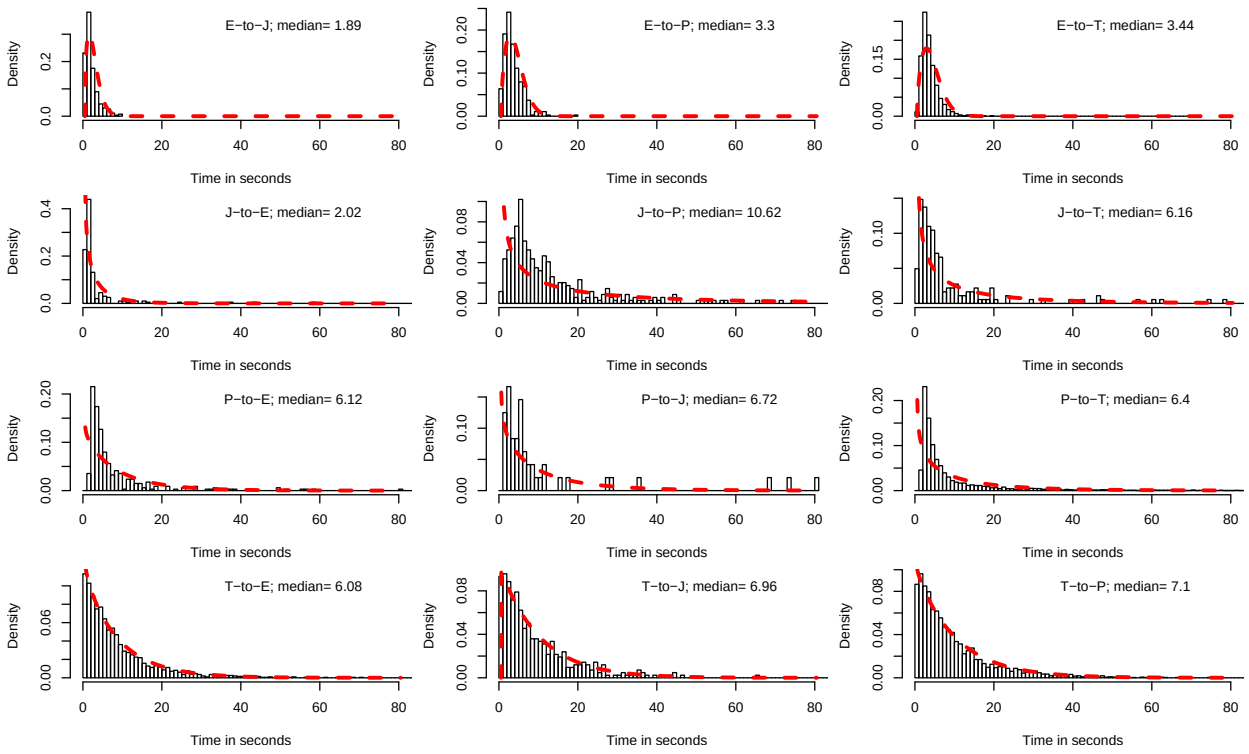

Figure 8: Estimated density distributions of sojourn time in the semi-Markov model compared to empirical data for Form 3. In each panel, the bars were produced by histogram, and the red dashed line by the semi-Markov model.

Table 11: Semi-Markov Model Fitting for Form 4. There were 241 students in the sample with essay keystroke data.

Label	Transition	Estimation	SD	LowerCI	UpperCI	H0	p.value
σ_{12}	E -> J	2.33	0.09	2.15	2.50	1.00	<0.0001
σ_{13}	E -> P	4.46	0.11	4.26	4.67	1.00	<0.0001
σ_{14}	E -> T	4.37	0.05	4.27	4.47	1.00	<0.0001
σ_{21}	J -> E	3.98	0.29	3.42	4.55	1.00	<0.0001
σ_{23}	J -> P	25.02	1.52	22.04	28.00	1.00	<0.0001
σ_{24}	J -> T	7.05	0.58	5.91	8.19	1.00	<0.0001
σ_{31}	P -> E	10.27	0.53	9.24	11.30	1.00	<0.0001
σ_{32}	P -> J	10.63	1.08	8.51	12.76	1.00	<0.0001
σ_{34}	P -> T	10.22	0.18	9.86	10.57	1.00	<0.0001
σ_{41}	T -> E	7.76	0.15	7.47	8.06	1.00	<0.0001
σ_{42}	T -> J	6.09	0.29	5.53	6.66	1.00	<0.0001
σ_{43}	T -> P	9.03	0.13	8.78	9.28	1.00	<0.0001
ν_{12}	E -> J	1.11	0.03	1.05	1.17	1.00	0.0004
ν_{13}	E -> P	1.55	0.04	1.48	1.63	1.00	<0.0001
ν_{14}	E -> T	1.58	0.02	1.55	1.62	1.00	<0.0001
ν_{21}	J -> E	0.64	0.02	0.61	0.68	1.00	<0.0001
ν_{23}	J -> P	0.66	0.02	0.63	0.69	1.00	<0.0001
ν_{24}	J -> T	0.51	0.01	0.49	0.53	1.00	<0.0001
ν_{31}	P -> E	0.78	0.02	0.74	0.81	1.00	<0.0001
ν_{32}	P -> J	0.95	0.06	0.84	1.07	1.00	0.4503
ν_{34}	P -> T	0.72	0.01	0.71	0.74	1.00	<0.0001
ν_{41}	T -> E	0.94	0.01	0.92	0.97	1.00	<0.0001
ν_{42}	T -> J	0.68	0.02	0.65	0.72	1.00	<0.0001
ν_{43}	T -> P	0.98	0.01	0.96	1.00	1.00	0.0208

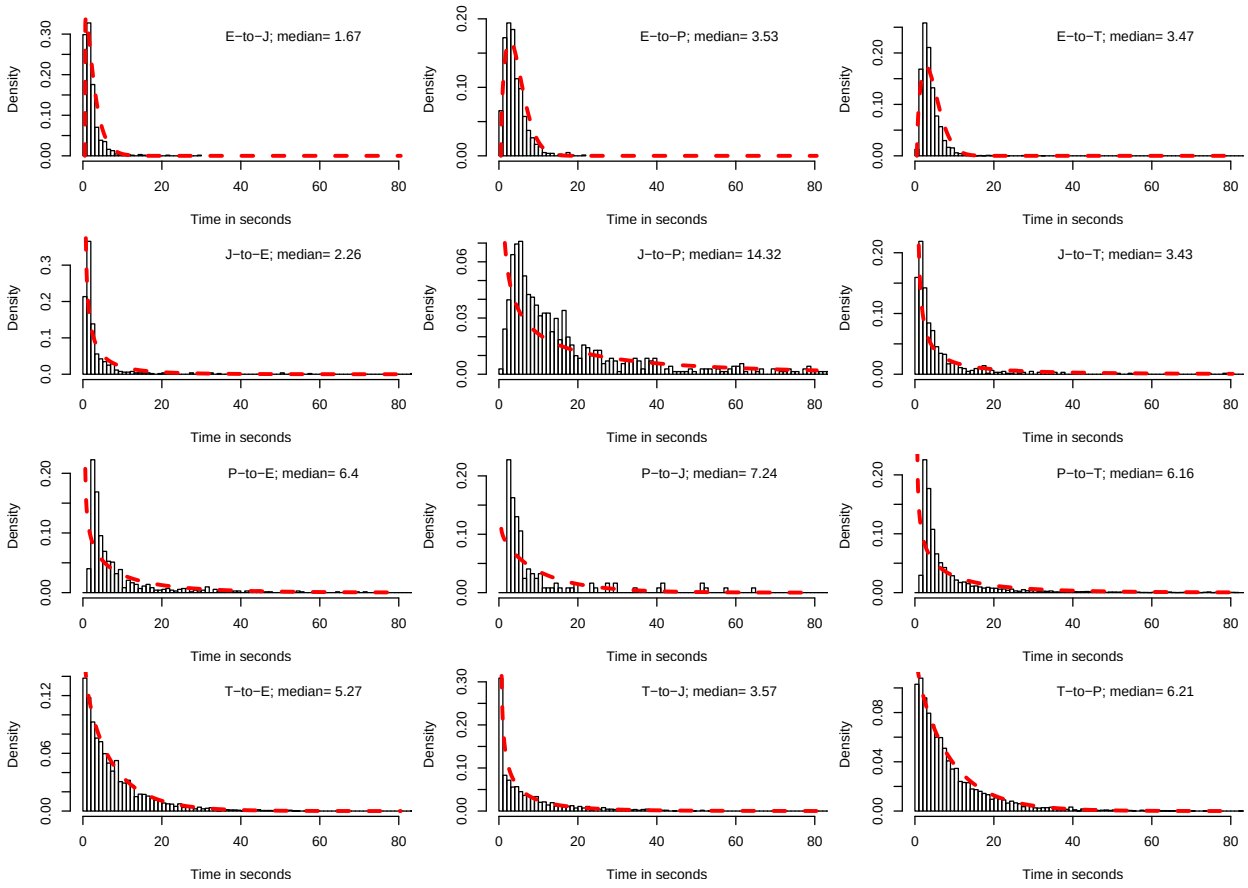

Figure 9: Estimated density distributions of sojourn time in the semi-Markov model compared to empirical data for Form 4. In each panel, the bars were produced by histogram, and the red dashed line by the semi-Markov model.