Chinese/English Journal of Educational Measurement and Evaluation | 教育测量与评估双语季刊

Volume 2 | Issue 1 Article 1

2021

How Do Educationally At-Risk Men and Women Differ in Their Essay-Writing Processes?

Randy E. Bennett
ETS

Mo Zhang
ETS

Sandip Sinharay
ETS

Recommended Citation

Bennett, Randy E.; Zhang, Mo; and Sinharay, Sandip (2021) "How Do Educationally At-Risk Men and Women Differ in Their Essay-Writing Processes?," *Chinese/English Journal of Educational Measurement and Evaluation | 教育测量与评估双语季刊*: Vol. 2 : Iss. 1 , Article 1.
Available at: https://www.ce-jeme.org/journal/vol2/iss1/1

How Do Educationally At-Risk Men and Women Differ in Their Essay-Writing Processes?

Randy E. Bennett [a], Mo Zhang [a], and Sandip Sinharay [a]

[a] Educational Testing Service

Abstract
This study examined differences in the composition processes used by educationally at-risk males and females who wrote essays as part of a high-school equivalency examination. Over 30,000 individuals were assessed, each taking one of 12 forms of the examination's language arts writing subtest in 23 US states. Writing processes were inferred using features extracted from keystroke logs and aggregated into seven composite indicators. Results showed that females earned higher essay and total language arts writing composite scores than did males, but only by trivial amounts. More pertinent was that, after controlling for language arts writing composite score, age, and essay prompt, all seven process indicators showed nontrivial, statistically significant differences, the most notable being for indicators related to fluency and different aspects of editing. The study's findings are consistent in important ways with those from other investigations of school-age students and adults, and with results from both online and paper-based writing tasks. Implications are offered for conducting similar research for individuals composing in character-based languages like Chinese.

Keywords
writing;
problem-solving processes;
performance assessment;
gender differences

It is well-established that girls perform better than boys on essay tests like those used by the US National Assessment of Educational Progress (NAEP). That gender difference has been observed on NAEP at medium effect sizes in the 8th and 12th grades from the 1988 through the 2011 writing assessment cycles (Reilly, Neumann, & Andrews, 2019). Such differences may be important because writing proficiency is related to success in US postsecondary education (Bridgeman & Lewis, 1994; Norris, Oppler, Kuang, Day, & Adams, 2006), and is often desired for employment.

With the transition of writing assessment from paper delivery to computer, we can go beyond observing differences in outcomes to explore the writing processes that produce an essay. These writing processes can be inferred from the keystrokes, time latencies, and other events captured as part of computer-based testing. Among other things, studying these processes may offer insight into how important demographic groups diverge in their approaches to composition.

In the current study, we examine gender differences in writing processes among a particular population segment, educationally at-risk adults. We consider these individuals to be at-risk because they did not complete their high school education and thus are more likely to have lower earnings and employment rates, coupled with higher poverty and incarceration rates (Breslow, 2012; Scott, Zhang, & Koball, 2015).

1 Literature Summary

Writing proficiency is commonly judged through an examination that includes one or more writing samples, sometimes supplemented with selected-response questions. These methods are employed for evaluating the proficiency of populations, as in the NAEP writing assessment, as well as the competency of individuals, as in the Test of English as a Foreign Language (TOEFL iBT) or the Chinese Gaokao (college entrance examination). Such assessments are often criticized because they measure writing skill in a timed, on-demand context quite different from the more

CONTACT: Randy E. Bennett. RBennett@ETS.org. MS R047, Educational Testing Service, Princeton, NJ, 08541

extended activity involved in producing a class paper. Even so, writing assessments not only predict success in college (Bridgeman & Lewis, 1994; Norris et al., 2006), but have demands not entirely different from the ones imposed by the world of work, where a piece of writing often needs to be quickly completed for a supervisor or client. Indeed, writing skills are among the most frequently indicated learning outcomes by US college and university officials (Hart Research Associates, 2016).

In test situations, research suggests that writing processes are associated with essay quality (Bennett, Zhang, Deane, & van Rijn, 2020; Deane & Zhang, 2015). Thus, the literature related to writing processes in test situations may be relevant to understanding the nature of outcome differences in writing among individuals as well as groups.

In assessment contexts, much of the research has utilized essays composed as part of middle-school English language arts assessments. Those studies have found a variety of process measures to be related to human judgments of essay quality. For example, fluency process features such as overall typing speed and average burst length (the number of characters typed consecutively before a pause), were positively related to quality, whereas within-word pauses and the average inter-key interval, were negatively related, being suggestive of difficulties in spelling or typing (Almond, Deane, Quinlan, Wagner, & Sydorenko, 2012; Bennett et al., 2020; Deane & Zhang, 2015; Guo, Deane, van Rijn, Zhang, & Bennett, 2018; Zhang, Hao, Li, & Deane, 2018). Also predictive of quality were measures of the extent of editing (Tate & Warschauer, 2019), indicators linked to effort like total time spent writing and the number of words started (Bennett et al., 2020; Deane & Zhang, 2015), and measures associated with planning, such as the extent of between-sentence pauses (Bennett et al., 2020; Deane & Zhang, 2015).

Not only has research shown process measures to relate to essay quality, but after controlling for quality, process measures have been found to differentiate demographic groups. In one study that asked students to write argumentative essays using evidence from source materials, low socio-economic-status (SES) students as well as Black students were less efficient than their comparison groups (high SES and White students, respectively), producing final texts that were smaller portions of the total number of keystrokes made (Guo, Zhang, Deane, & Bennett, 2019). Most pertinent to the current investigation, however, was that females were more fluent than males with similar essay scores. Females typed faster, used more complex words, spent longer time in text production, and engaged in quick and frequent editing and pauses. In a second study that also used writing from source materials and conditioned on essay score, females were again more fluent, edited more, and appeared to need to pause less in locations associated with planning (Zhang, Bennett, Deane, & van Rijn, 2019).

Cognitive writing theory and research offer insight into some of the above results. This work suggests that lower- and higher-level processes compete for limited working memory (Kellogg, 2001; McCutchen, 1996, 2011). When lower-level processes like transforming ideas into words and sentences, typing, and spelling are automated, a writer can give more attention to the higher-order planning and revision processes needed to generate quality text. In fact, research indicates that, compared to beginners, accomplished writers produce text fluently in relatively long chunks and bursts; spend more time planning, generating text, and revising; and tend to stop at natural planning junctures like clause and sentence boundaries (Connelly, Dockrell, Walter, & Critten, 2012; Kaufer, Hayes, & Flower, 1986; McCutchen, 2006; McCutchen, Covill, Hoyne, & Mildes, 1994). Cognitive research using keystroke logging has generally reaffirmed these findings (Baaijen, Galbraith, & De Glopper, 2012; Leijten & Van Waes, 2013; Strömqvist, Holmqvist, Johansson, Karlsson, & Wengelin, 2006; Sullivan & Lindgren, 2006; Wengelin, 2006).

In sum, writing assessments measure skills related to success in higher education and that are arguably important for the workplace. On such essay examinations, US girls have consistently performed at higher levels than boys in both middle school and high school. Studies of middle-school girls and boys with similar essay scores indicate that these groups use notably different composition processes when writing persuasive essays based on source materials.

The goal of the current study was to determine whether such process differences extend to educationally at-risk adults. Two research questions were addressed:

1. Are there differences in writing processes between educationally at-risk adult males and females? More specifically, do the groups differ in such processes as fluency, editing, or other behaviors that cognitive theory and research have suggested are implicated in producing quality writing?
2. If so, are those process differences simply a function of disparities in writing-skill level? That is, do group-related process differences persist after

accounting for other factors, including writing quality.

2 Method

2.1 Participants

Participants came from a testing program that is administered on paper and by computer for awarding equivalency diplomas to individuals who did not complete their high school education. For the period September 2017 through August 2018, we selected all examinees who were administered the language arts writing subtest on computer. Twelve parallel forms of that subtest were administered during this period, resulting in 32,164 records from testing centers in 23 US states. After removing the subsequent records of examinees who repeated the exam, records with corrupted log files, and records for which process evaluation would not be meaningful (zero essay score indicating a blank or off-topic response), 30,788 unique examinees remained.

For analysis purposes, the sample was randomly split into main and replication groups. The demographic composition of the main and replication groups is given in Table 1. As can be seen, there was a somewhat larger percentage of males than females (53% to 47%). The mean age was 24.5, with a range extending from the teen years through senior citizenship. About half the participants were White, with the next largest ethnic groups being Black examinees (16%) and Hispanic examinees (15% and 16% in the main and replication samples, respectively). Over 90% of participants indicated that they communicated in English best. With respect to highest education level, about a third of the participants said this level was 10th grade or lower, a similar proportion selected 11th or 12th grade, and a third did not provide a value. For mother's highest education level, the plurality (∼36%) did not report, close to a fifth indicated high school completion, and another fifth said that parent had some college or a degree. Finally, about a third of participants indicated they were working full- or part-time, just over a third reported being unemployed or not in the labor force, and slightly less than this proportion did not report.

In Table 2 are the demographic distributions by gender. As can be seen, the total male and female samples were generally similar, with the largest differences on most variables being greater percentages of missing data for males (e.g., for ethnicity, highest education level, mother's highest education, employment status). These missing data suggest the need to view with caution the small differences observed on some of the reported characteristics for those same variables (e.g., females appeared more likely to be White but males had more missing data for this question). An exception was age, due to a requirement to supply birthdate, with females on average being older than males by about two years.

2.2 Instrument

Each parallel form of the high school equivalency examination included five subtests. Subtests could be taken on the same day or scheduled on different days in any order depending on the examinee's preference. The language arts reading subtest had 40 (operational) multiple-choice questions assessing the ability to understand, comprehend, interpret, and analyze a variety of reading material. The language arts writing subtest contained 50 multiple-choice items and one essay prompt (i.e., question) measuring the ability to edit and revise written text, and to generate and organize ideas in writing. The third subtest, mathematics, was comprised of 50 multiple-choice questions focusing on fundamental concepts and reasoning skills. The science subtest included 50 multiple-choice questions tapping proficiency in science content knowledge, applying principles of scientific inquiry, and interpreting and evaluating scientific information. Finally, the social studies subtest contained 50 multiple-choice questions designed to measure skill in analyzing and evaluating domain-relevant information.

For each subtest, scale scores ranged from 1-20. For the essay prompt, which was included as part of the language arts writing composite score, a raw score was also reported on a 0-6 scale. The passing criteria used by most jurisdictions are a score of at least 8 on each of the five subtests, a score of at least 2 on the essay question, and a total score of 45 or higher (the sum of the five subtest scale scores).

The essay prompt, which was the focus of this study, called for the examinee to read two presented source passages that give differing perspectives on an issue (e.g., whether success is more the result of talent or of hard work). The two passages typically had 750-800 words combined and were more or less equivalent in length. The examinee must then compose an essay that gives and explains his or her opinion, using evidence from the sources.

2.3 Scoring

Essay scoring was completed by professional raters as part of the program's operational scoring process. Raters

Table 1
Demographic Distributions for Study Samples

Background variable	Main sample		Replication sample	
	(*N* = 15,368)		(*N* = 15,420)	
Age	24.5	(*SD*=9.1)	24.5	(*SD*=9.1)
Gender				
Female	7193	(47%)	7231	(47%)
Male	8175	(53%)	8189	(53%)
Ethnicity				
White	7527	(50%)	7421	(48%)
Asian	221	(1%)	199	(1%)
Black or African American	2425	(16%)	2412	(16%)
American Indian or Alaskan Native	258	(2%)	254	(2%)
Native Hawaiian or Pacific Islander	40	(0.3%)	32	(0.2%)
Hispanic	2336	(15%)	2419	(16%)
Multiracial	644	(4%)	633	(4%)
Other race	231	(2%)	230	(2%)
Prefer not to respond	1686	(11%)	1820	(12%)
Communicate best in English				
Yes	14,310	(93%)	14,343	(93%)
No	974	(6%)	991	(6%)
Missing	84	(0.6%)	86	(0.6%)
Highest education level				
Below 9th	786	(5%)	764	(5%)
9th	1500	(10%)	1567	(10%)
10th	2878	(19%)	2868	(19%)
11th	4013	(26%)	3998	(26%)
12th	1116	(7%)	1098	(7%)
Missing	5075	(33%)	5125	(33%)
Mother's highest education				
Grade school	495	(3%)	526	(3%)
Some high school	1455	(9%)	1477	(10%)
Completed high school	2858	(19%)	2831	(18%)
Some college	1705	(11%)	1594	(10%)
Associate degree	524	(3%)	566	(4%)
Bachelor's degree	735	(5%)	750	(5%)
Postgraduate education	478	(3%)	446	(3%)
Unknown	1563	(10%)	1570	(10%)
Missing	5555	(36%)	5660	(37%)
Employment status				
Part time	2248	(15%)	2298	(15%)
Full time	2757	(18%)	2736	(18%)
Unemployed	3750	(24%)	3682	(24%)
Not in labor force	2030	(13%)	2014	(13%)
Missing	4583	(30%)	4690	(30%)

Table 2
Demographic Distributions by Gender

Background variable		Males		Females	
		(*N* = 16,364)		(*N* = 14,424)	
Age		23.5	(*SD*=8.7)	25.6	(*SD*=9.4)
Ethnicity					
	White	7856	(48%)	7092	(49%)
	Asian	185	(1%)	235	(2%)
	Black or African American	2505	(15%)	2332	(16%)
	American Indian or Native Alaskan	270	(2%)	242	(2%)
	Native Hawaiian or Pacific Islander	33	(0.2%)	39	(0.3%)
	Hispanic	2358	(14%)	2397	(17%)
	Multiracial	699	(4%)	578	(4%)
	Other race	256	(2%)	205	(1%)
	Prefer not to respond	2202	(13%)	1304	(9%)
Communicate best in English					
	Yes	15,258	(93%)	13,395	(93%)
	No	1001	(6%)	964	(7%)
	Missing	105	(0.6%)	65	(0.5%)
Highest education level					
	Below 9th	694	(4%)	856	(6%)
	9th	1403	(9%)	1664	(12%)
	10th	2757	(17%)	2989	(21%)
	11th	4162	(25%)	3849	(27%)
	12th	1167	(7%)	1047	(7%)
	Missing	6181	(38%)	4019	(28%)
Mother's highest education					
	Grade school	357	(2%)	664	(5%)
	Some high school	1193	(7%)	1739	(12%)
	Completed high school	2814	(17%)	2857	(20%)
	Some college	1606	(10%)	1693	(12%)
	Associate degree	543	(3%)	547	(4%)
	Bachelor's degree	830	(5%)	655	(5%)
	Postgraduate education	496	(3%)	428	(3%)
	Unknown	1722	(11%)	1411	(10%)
	Missing	6803	(42%)	4412	(31%)
Employment status					
	Part time	2099	(13%)	2447	(17%)
	Full time	2860	(17%)	2633	(18%)
	Unemployed	3863	(24%)	3569	(25%)
	Not in labor force	1882	(12%)	2162	(15%)
	Missing	5660	(35%)	3613	(25%)

were trained with a released prompt, certified as ready for scoring based on their grading of responses to that prompt, and then calibrated each day on an operational prompt. Their operational scoring performance was monitored by a leader who intervened if the rater drifted from the pre-assigned scores given by experts to validity papers seeded into the rater's queue.

The scoring rubric covered four content categories (development of a central position or claim, organization of ideas, language facility, and writing conventions), with an essay to be rated holistically taking those categories into consideration. The rubric for a given essay question was accompanied by prompt-specific notes and benchmarks that showed the range of responses for each score point.

Each response was scored by two raters. If the two raters disagreed by more than one point, the response was sent for additional grading by a third rater.

In our dataset, after excluding essays receiving zero scores, the median exact agreement between the first and second raters taken over the 12 essay prompts was 78% (prompt range = 76% to 79%) and the median quadratic weighted kappa values were .65 (prompt range = .61 to .68). For our study, the mean rating taken across the first two raters was employed.

From each examinee's essay, 78 process features were extracted, taken from a larger set of several hundred that have been developed through a research program on writing process analysis (Deane, 2014; Deane & Zhang, 2015). The 78 features included counts (e.g., total number of keystrokes), durations (e.g., minimum duration of continued backspace events), rates (e.g., maximum rate of long-distance jump-to-edit actions), and probabilities (e.g., relative probability of a word in the final text ever having been misspelled). These features were factor analyzed using exploratory methods to identify how they might be organized into process scales. Twelve principal-components analyses with promax oblique rotation were conducted, one analysis for each prompt. Features were retained that consistently loaded $>.30$ in the same direction on the same factor for at least 8 of the 12 analyses, a decision rule intended to produce a feature set that behaved reasonably consistently across prompts. From the results, a seven-factor solution that accounted for approximately 62% of the variance was selected based on interpretability and parsimony. This solution included 59 features, of which 51 (86%) met the selection criteria for either 11 or 12 prompts, five for ten prompts, two for nine prompts, and one for eight prompts.

The above procedure was used to create seven process scales, with the features for each scale corresponding to the features that consistently loaded on the related factor. For each of the scales, a composite score was created for each examinee by standardizing each feature score to a mean of 0 and a standard deviation of 1, and then summing the results to create a process indicator. Because these process indicators were not composed of equal numbers of features, they cannot be directly compared to one another.

The process indicators, their associated features, and the valence of each feature as determined from the factor analysis are given in the Appendix. Each indicator was named based on the preponderance of features that composed it. The first indicator, Between-Word Speed, was composed of nine features such as the rate of spacing keystrokes between words (positive valence) and pause durations (negative valence). This indicator might be thought of as an index of facility in transitioning to the next word. Faster transitions imply some combination of greater command of lexical retrieval (i.e., word finding), syntactic encoding (i.e., arranging words in appropriate order), and typing facility. The second indicator was Large-Burst Fluency, which was composed of 11 features. This indicator combined various measures of burst length for long segments of text (positive valence) with such indicators of typing precision as the probability of a word in the final text ever having been misspelled (negative valence). Together, these features indexed automaticity for generating relatively long bursts without error, corrected or not. Third, Productivity (eight features) reflected the amount of text produced (though not necessarily retained in the final essay submission). It included such features as the total number of keystrokes in the session (positive valence) and the probability of time being spent rearranging text (negative valence to account for the inefficiency implied by a high probability). As such, Productivity is suggestive of fluency and efficiency. Fourth, Deletion Editing (nine features) was composed primarily of variables related to cutting text (e.g., time spent, number of characters cut, maximum rate of long-distance jumps), all of which had positive valence. This indicator might be taken as indexing the propensity to make many quick cuts, including ones away from the current cursor position. The fifth indicator was Jump Editing, defined as a change to text (insertion or deletion) requiring moving from the current cursor position to a nonadjacent location within the current word, elsewhere in the same sentence, or beyond. This indicator had 12 features, such as the time spent on jump edits and the maximum duration in keystrokes, all having positive valence. Jump Editing suggests text-monitoring behavior at the word, phrase, sentence, and occasionally whole-text level. Sixth, was Between-Sentence Transition and Backspace Speed (six features). Such features as the maximum rate of between-sentence space keystrokes (positive valence) and the minimum duration of continued backspaces (negative valence) comprised it. This fluency-related indicator appeared to primarily tap the degree of flow from one sentence to the next, along with the extent of rapid text removal through backspacing of something already written. Finally, the four Paragraphing features, all with positive valence, concerned the number and duration of keystrokes related to line breaks. This indicator might be taken as related to the extent of effort devoted to organization and to planning in between paragraphs.

2.4 Data Analysis

Participants were randomly assigned to one of two samples, the first designated as "main" and the other as "replication." All analyses conducted in the main sample were also run in the replication sample, and a result was considered meaningful only if it was statistically significant in both samples.

We conducted a multivariate analyses of variance (MANOVA) to determine whether the study groups differed in their writing processes. Gender was the independent variable and the scores on each of the process indicators were the dependent variables. Univariate tests were conducted to detect whether the dependent variables were significantly different statistically between the gender groups. Next, a MANCOVA was run with the same independent and dependent variables but with the addition of language arts writing composite score, essay prompt (dummy coded), and age as covariates (as well as with all two-way interactions). This analysis was run to determine if the observed process differences were simply a reflection of disparities between the gender groups on these variables (i.e., in writing skill, age, or in the prompts assigned).

All analyses were conducted using the statistical software SAS. The proc factor procedure was used for factor analyses, and the proc glm procedure was used for MANOVA/MANCOVA (Pett, Lackey, & Sullivan, 2003; Kuehl, 2000). For all analyses, statistical tests were two-tailed, with .05 used as the α level in most situations. For the MANOVA/MANCOVA univariate follow-up tests, the Bonferroni correction was applied, with $\alpha/7$ (the number of process indicators) = .0071. The correction was applied to these tests because the tests directly addressed the study's research questions.

3 Results

3.1 Summary Statistics

Table 3 gives the mean high-school-equivalency subtest scores. There were no statistically significant differences between the groups on the language arts reading subtest, but females scored statistically significantly higher than males on the language arts writing subtest. These differences were, however, trivial in Cohen's (1988) categorization (i.e., < 0.20 *SD* units). In contrast, males achieved statistically significantly higher scores than females on the remaining three subtests—mathematics, science, and social studies—with the differences being relatively small in all cases (i.e., between 0.22 and 0.36 *SD* units).

Table 3
Mean High-School-Equivalency Examination Scores by Gender

	Main sample		Replication sample	
	Male	Female	Male	Female
	Language arts writing (scale=1-20)			
Mean	11.40	11.80**	11.30	11.60**
SD	3.39	3.27	3.36	3.30
	Language arts reading (scale=1-20)			
Mean	11.80	11.80	11.70	11.70
SD	4.18	4.12	4.22	4.15
	Mathematics (scale=1-20)			
Mean	11.90	10.60**	11.80	10.50**
SD	4.54	4.63	4.59	4.62
	Science (scale=1-20)			
Mean	13.90	13.00**	13.80	12.90**
SD	4.00	3.90	4.04	3.87
	Social studies (scale=1-20)			
Mean	13.70	12.20**	13.70	12.10**
SD	4.49	4.50	4.44	4.51

Note. *N* range = 6,253 to 8,289 for gender group within sample.
**p <.0001.

Table 4 gives the mean essay scores, lengths, and times. The differences between the gender groups were statistically significantly different for all three measures. Females scored higher than males, wrote longer essays, and spent more time. In all cases, the differences were trivial.[1]

Table 4
Mean Essay Scores, Lengths, and Times by Gender

	Main sample		Replication sample	
	Male	Female	Male	Female
N	8175	7193	8189	7231
		Essay score (scale=1-6)		
Mean	2.60	2.80^{**}	2.60	2.70^{**}
SD	0.89	0.86	0.89	0.87
		Essay length (in words)		
Mean	288.00	308.00^{**}	289.00	301.00^{**}
SD	130.00	131.00	131.00	129.00
		Time on essay task (in minutes)		
Mean	33.70	34.50^{*}	33.50	34.30^{*}
SD	16.40	15.90	16.20	16.10

$^{*}p<.005$.$^{**}p<.0001$.

Table 5 gives the inter-correlations among the measures and the internal consistency reliability coefficients (where such indices could be calculated or were otherwise available). The table includes values from both main and replications samples, with almost all results statistically significant and only marginally different across samples. As the table shows, the process indicators were quite reliable (alpha range = .87 to .96). In addition, the indicators generally had low correlations with one another (r range $= |.01|$ to $|.46|$ in the main sample and $|.01|$ to $|.45|$ in the replication sample), attesting to the relative independence of the process measures.

The correlations for Productivity were among the largest. This indicator was understandably related to fluency indicators like Between-Word Speed (r = .42) and Between-Sentence Transition and Backspace Speed (r = .41), because less time spent on these transitions leaves more time for generating text. Productivity was also related to the editing indicators, in the mid-.40s with Jump Editing and .37 with Deletion Editing, presumably because high productivity provides more text to edit. This supposition is consistent with the indicator's strong relation to essay length (r = .83). Finally, Productivity was correlated in the mid-.50s with time on essay task, .68 with essay score and .42 with language arts writing composite score.

Additionally of note in Table 5 is that all process indicators were positively related to essay score and to language arts writing composite score. Including Productivity, the correlations with essay score ranged from .17 to .68, and with language arts writing composite score from .19 to .42. The indicator correlating next most highly with these two quality measures was Between-Word Speed, relating to essay score at .38 and language arts writing composite score at .37. Also having notable relationships with essay score were Jump Editing (r = .34) and Paragraphing (r = .30). The lowest correlations with essay score (r = .17), and second lowest with language arts writing composite score (r = .21 for the main sample and .20 for the replication sample), were observed for Large-Burst Fluency. This indicator also had very small (and always negative) correlations with the other indicators, as well as a small negative correlation with time on task, suggesting a slight tendency for students who generated long text segments precisely to use less time overall.

Finally, as might be expected, the process indicators appeared to be more highly related to language arts writing test score than to mathematics score. Though generally low, their relations to mathematics score were, however, often at levels quite similar to those for the social studies, science, and reading test scores. This result might reflect the influence of a general fluency factor common to the measures. In fact, the highest observed correlations with mathematics test score were for Between-Word Speed and Productivity, both of which are fluency related.

3.2 Research Question 1

To examine whether there were differences in writing processes between gender groups, we compared the mean scores on each process indicator across groups (see Table 6). Results indicated a statistically significant overall effect for gender, with the MANOVA test statistic Pillai's Trace = 0.045, $F(7, 15360) = 103.77$, $p <.0001$ for the main sample and Pillai's Trace = 0.043, $F(7, 15412) = 98.09$, $p <.0001$ for the replication sample. The univariate results showed that the male and female means significantly differed statistically for all indicators ($p <.0001$) except Large-Burst Fluency (p = .2710). The results indicate that the female group was faster in transitions between words, generated more text, engaged in more deletion editing and jump editing, was faster in making between-sentence transitions and in backspacing, and devoted more effort to organization and possibly planning in between paragraphs.

[1] Similarly trivial group differences were found for the multiple-choice portion of the language arts writing subtest, suggesting no noteworthy relationship between the size of the gender difference and item type.

Table 5

Correlations Among Study Variables

	Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1.	Between-word speed	**.96**	-.03	.42	.25	.07	.34	.10	.38	.37	-.15	.35	.28	.24	.22	.27
2.	Large-burst fluency	-.01	**.94**	-.03	-.12	-.15	-.10	-.02	.17	.20	-.22	.14	.18	.10	.15	.15
3.	Productivity	.42	-.01	**.94**	.37	.45	.41	.30	.68	.42	.55	.83	.30	.26	.24	.27
4.	Deletion editing	.24	-.12	.37	**.93**	.32	.21	.17	.25	.25	.31	.22	.20	.17	.15	.17
5.	Jump editing	.08	-.14	.46	.33	**.87**	.20	.28	.33	.19	.59	.36	.14	.09	.10	.11
6.	Between-sentence transition & backspace speed	.33	-.10	.41	.20	.21	**.87**	.17	.24	.22	.22	.21	.18	.15	.14	.15
7.	Paragraphing	.11	-.03	.28	.17	.30	.18	**.94**	.30	.21	.27	.26	.17	.14	.15	.16
8.	Essay score	.38	.17	.68	.25	.34	.24	.30	–	.60	.31	.74	.42	.35	.36	.38
9.	Language arts writing	.37	.21	.42	.24	.20	.20	.20	.60	**.71**	.07	.40	.71	.54	.65	.67
10.	Time on essay task	-.16	-.22	.54	.31	.58	.22	.26	.30	.06	–	.42	.08	.05	.07	.05
11.	Essay length	.34	.16	.83	.22	.37	.21	.24	.73	.40	.41	–	.27	.26	.23	.26
12.	Language arts reading	.28	.16	.31	.18	.16	.17	.15	.42	.70	.09	.27	**.85**	.50	.70	.69
13.	Mathematics	.25	.10	.25	.17	.08	.13	.13	.34	.53	.04	.24	.50	**.75**	.53	.61
14.	Social studies	.22	.15	.24	.16	.11	.13	.14	.36	.65	.07	.22	.69	.53	**.83**	.70
15.	Science	.26	.15	.27	.18	.14	.15	.16	.38	.67	.07	.25	.69	.61	.70	**.85**

Note. In bold on the diagonal are coefficient alpha values for process indicators estimated from the current data and for equivalency examination section scores described in the test manual as estimated from the 2015 administration. Main sample values are below diagonal and replication sample values are above diagonal. Values of |.07| or above for main sample, and |.04| or above for replication sample, are statistically significantly different from zero at p <.0001. Main sample N range = 13,557 to 15,368; replication sample N range = 13,618 to 15,420.

3.3 Research Question 2

Because group differences in writing processes may simply reflect disparities in other variables, we compared group performance on the process indicators after including language arts writing composite score, age, and prompt as covariates. This analysis produced a statistically significant overall effect for the main sample, Pillai's Trace = 0.055, $F(7,\ 15347) = 128.36$, $p <.0001$, and for the replication sample, Pillai's Trace = 0.052, $F(7,\ 15399) = 121.79$, $p <.0001$[2,3]. In contrast to the MANOVA model, here the univariate tests were statistically significant for gender on all seven process indicators, with Large-Burst Fluency changing its status.

The adjusted indicator means and effect sizes are shown in Table 7. Of note is that the covariance adjustments generally had minimal impact. More importantly, examination of the effect sizes suggests that gender differences on four process indicators are worth noticing: Between-Word Speed ($d = 0.39$ in the main sample, 0.35 in the replication sample), Productivity ($d = 0.27$ and 0.23), Deletion Editing ($d = 0.30$ and 0.33), and Jump Editing ($d = 0.28$ and 0.26). These effects would be characterized as small in Cohen's (1988) classification (i.e., 0.20 to 0.49). The remaining three indicators had effects of 0.14 or less, which is of trivial magnitude.

4 Discussion

This study examined differences in the processes used by educationally at-risk males and females in composing essays for a high-school equivalency examination. The study sample included over 30,000 individuals, each taking one of 12 forms of the examination's language arts writing subtest, with each form containing one essay task. Examinees came from almost half of the US states. Writing processes were inferred using features extracted from keystroke logs and aggregated into seven composite indictors, with results considered meaningful only if replicated across randomly parallel samples. We found females to earn higher essay and language arts writing composite scores than males, but only by trivial amounts (i.e., <0.20 SD units). More interesting was that males and females differed statistically significantly on six of the seven process indicators. After controlling for language arts writing composite score, age, and essay prompt, all seven indicators showed statistically significant differences, with the most notable effect sizes being for Between-Word Speed, Productivity, Deletion Editing, and Jump Editing.

That educationally at-risk adult females received only marginally higher essay and language arts writing composite scores than males is somewhat surprising. From NAEP, we know that there have been medium-size differences in writing performance at both the 8th and 12th grade levels across many assessment cycles (Reilly et al., 2019). Aside from differences in the way in which writing was assessed in the high-school equivalency and NAEP instruments, a major difference is in the examinee populations. NAEP tests nationally representative samples of young adults, whereas our samples are from a lower segment of the proficiency distribution that is on average several years older as well as considerably more variable in age. From NAEP, we also know that gender differences in writing at the lower end of the distribution are larger (Reilly et al., 2019), which raises the possibility that different segments of the male and female populations are opting to take equivalency examinations generally. Males, for example, are several times more likely than females to be incarcerated (Glaze & Kaeble, 2014, p. 6; IES, 2014) and the incarcerated are on average less literate than the general population (IES, 2014). Even though equivalency examinations are offered in many prisons, one might speculate that such assessments are less available to prisoners than to the general population, thereby removing more lower-skilled males than females from the assessed cohort.

A second factor that might be impacting our results is that equivalency examination policies vary by state. That is, some states have a single examination (e.g., Missouri) whereas other states allow choice among two or more examinations (e.g., California). In addition, some states allow examinees to choose the testing mode, computer or paper. It is possible that, given a choice, females and males differentially select examinations and/or delivery modes. When we compared states offering the study instrument as the only equivalency test with states that allowed a choice of more than one equivalency exam, indeed the gender differences for essay and language arts writing composite scores were noticeably larger in the former

[2] Consistent statistically significant main effects were observed for language arts writing score and for age on all indicators; and for prompt on Between-Word Speed, Large-Burst Fluency, Productivity, and Deletion Editing.

[3] When the same model was run including all two-way interactions, those interactions were not consistently statistically significant so only the results of the main-effects model are reported here.

Table 6

Observed Process Indicator Means (SDs) by Gender

	Main sample			Replication sample		
Gender	Mean	*SD*	Range	Mean	*SD*	Range
			Between-word speed			
Female	1.08	7.33	-25.12 to 23.52	0.89	7.39	-24.76 to 22.81
Male	−0.95*	8.13	-25.13 to 23.52	−0.79*	8.10	-24.76 to 23.52
			Large-burst fluency			
Female	−0.07	8.69	-19.88 to 36.41	−0.08	8.72	-19.89 to 36.33
Male	0.06	8.92	-19.49 to 36.26	0.07	8.80	-19.35 to 35.12
			Productivity			
Female	0.98	6.77	-13.40 to 26.04	0.78	6.75	-13.29 to 26.43
Male	−0.86*	6.54	-13.40 to 26.04	−0.68*	6.58	-13.29 to 26.43
			Deletion editing			
Female	1.20	7.56	-6.64 to 30.69	1.28	7.65	-6.61 to 30.38
Male	−1.06*	6.75	-6.64 to 27.07	−1.13*	6.60	-6.61 to 27.36
			Jump editing			
Female	1.24	7.37	-29.24 to 26.99	1.14	7.56	-29.77 to 27.15
Male	−1.09*	7.74	-29.44 to 24.10	−1.00*	7.56	-29.83 to 28.86
			Between-sentence transition & backspace speed			
Female	0.34	4.68	-9.52 to 15.78	0.17	4.65	-9.44 to 16.06
Male	−0.30*	4.71	-9.52 to 15.78	−0.15*	4.73	-9.44 to 16.06
			Paragraphing			
Female	0.22	3.49	-9.80 to 5.40	0.17	3.57	-9.48 to 5.42
Male	−0.19*	3.84	-9.80 to 5.40	−0.15*	3.81	-9.48 to 5.42

Note. Main sample N = 7,193 female and 8,175 male. Replication sample N = 7,231 female and 8,189 male. For each sample, results are collapsed across the same 12 prompts.

*p <.0071 for the univariate F-test difference between male and female means after Bonferroni correction.

group.[4] Moreover, although the pattern of practically important gender differences on the process indicators did not change, the differences in individual process indicator effects were on average larger in the group that offered only the study instrument, suggesting the possibility that our full-sample results might underestimate the magnitude of gender differences in educationally at-risk adults.[5] Such was particularly the case for Between-Word Speed and Productivity, which had effect sizes of 0.17 and 0.19, respectively, among multiple-instruments states and 0.31 and 0.31 in the study-instrument-only states.

What is the nature of the writing process differences that were detected? Two process indicators showing notable differences were associated, directly or indirectly, with fluency—Between-Word Speed, because it measured rapidity in moving to the next word, and Productivity, as a direct consequence of being able to rapidly generate and enter a significant amount of text. These indicators are of consequence as they relate both to essay score (Between-Word Speed $r = .38$; Productivity $r = .68$), and to language arts writing composite score (Between-Word Speed $r = .37$; Productivity $r = .42$). The two other process indicators showing prominent gender differences were

[4] The means of the state effect sizes (weighted by state N) in the multiple-instruments group ($N = 13,650$) and in the study-instrument-only group ($N = 17,135$) were 0.09 and 0.19 respectively for essay score. For language arts writing composite scores, the comparable means were 0.02 and 0.17.

[5] For the process indicator scores, the mean effect sizes in the multiple-instruments group and in the study-instrument-only group were for Between-Word Speed 0.17 and 0.31, Large-Burst Fluency 0.01 and 0.01, Productivity 0.19 and 0.31, Deletion Editing 0.31 and 0.36, Jump Editing 0.29 and 0.30, Between-Sentence Transition 0.07 and 0.13, and Paragraphing 0.10 and 0.11.

Table 7

Adjusted Process-Indicator Means and Effect Sizes Controlling for Age, Language Arts Writing Score, and Prompt for Males and Females in Main and Replication Samples

	Main sample		Replication sample	
Gender	Mean	Effect size	Mean	Effect size
	Between-word speed			
Female	1.17		1.05	
Male	−1.13*	0.39	−1.03*	0.35
	Large-burst fluency			
Female	−0.34		−0.20	
Male	0.25*	−0.08	0.33*	−0.08
	Productivity			
Female	0.89		0.70	
Male	−0.88*	0.27	−0.81*	0.23
	Deletion editing			
Female	1.16		1.17	
Male	−1.01*	0.30	−1.21*	0.33
	Jump editing			
Female	1.13		1.02	
Male	−0.95*	0.28	−0.92*	0.26
	Between-sentence transition & backspace speed			
Female	0.32		0.17	
Male	−0.33*	0.14	−0.23*	0.08
	Paragraphing			
Female	0.18		0.11	
Male	−0.12*	0.08	−0.11*	0.06

Note. Main sample N = 7,193 female and 8,175 male. Replication sample N = 7,231 female and 8,189 male. For each sample, results are collapsed across the same 12 prompts. Effect size is female mean minus male mean divided by the within-sample pooled standard deviation.

*p <.0071 for the univariate F-test difference between male and female means after Bonferroni correction.

connected to monitoring and revision—Deletion Editing and Jump Editing. These indicators also correlated with writing proficiency, though at lower levels (r range = .19 to .34).

The detected process differences were not only related to writing quality but very similar to ones found in middle-school students also taking an assessment requiring online writing from sources. Guo et al. (2019) found that, controlling for essay score, females typed faster, spent longer time in text production, and engaged in quick and frequent editing and pauses. In a different examinee sample, Zhang, Bennett, et al. (2019) reported that females were more fluent, edited more, and appeared to need to pause less in locations commonly associated with planning, also indicative of their higher fluency. Finally, earlier studies using simpler, paper-based writing tasks have documented comparable gender disparities from the elementary years to adulthood in productivity and fluency (Camarata & Woodcock, 2006; Jewell & Malecki, 2005; Malecki & Jewell, 2003). Our results, then, add to the evidence suggesting that gender differences in composition processes generalize to writing on computer and to at-risk adults taking an equivalency examination.

What are the implications of such process differences for achievement and instruction? A potentially important implication relates to the nature of writing processes. As found here and elsewhere (Bennett et al., 2020), greater levels of fluency and productivity are associated with higher essay performance. In theory, greater fluency, or automaticity for basic processes like word retrieval and typing, allows cognitive resources to be devoted to such higher-order concerns as idea generation and sentence planning (McCutchen et al., 1994; McCutchen, 1996). Additionally, the extent of editing has been associated with the quality of text (Tate & Warschauer, 2019; Zhang, Zhu, Deane, & Guo, 2019). That we observed notable gender disparities in fluency- and editing-related indicators after controlling for writing skill suggests that males can compensate for their disadvantages in these processes. In the current study, males were marginally lower in Between-Sentence Transition and Backspace Speed, suggesting they might be taking more time at these boundary locations to generate what to say next. In studies of middle-school students, a similar phenomenon has been observed with respect to pause time between bursts of text production, including sentences (Zhang, Bennett, et al., 2019); such pause times have been found in general to reflect planning behavior (e.g., Révész, Michel, & Lee, 2019). Females, because of their fluency advantage, might have had less need to pause at these locations, planning more effectively as they compose. In any event, identifying the compensatory mechanisms males use might help lower-scoring students to improve their writing. Among other methods, this identification might be facilitated by asking males to watch a replay of their composition process and describe what they were thinking and doing.

That females have evidenced a fluency advantage when writing online as well as on paper suggests that typing competency on its own is probably not responsible. A more likely explanation is somewhat higher verbal facility (Halpern, 2012), which itself might be caused by a combination of biological, social, and psychological factors (Reilly et al., 2019). In future research, the relative contribution of verbal fluency vs. typing skill might be evaluated by administering both an essay task and a simple retyping (copying) task (Deane et al., 2018), or by computing the speed with which easily recalled, high-frequency words are input (Zhang, Deane, Feng, & Guo, 2019).

With respect to research implications, it would be interesting to explore the extent to which group differences in writing process extend to other languages, including such character-based ones as Chinese. Among Chinese children, research suggests that both sentence-level handwriting fluency and gender are related to overall writing quality as measured on paper (Yan et al., 2012). Evaluating writing fluency in an online environment, however, poses special challenges because Chinese characters are not mapped directly to the computer keyboard in the same way that Roman letters are. In the most commonly used input method, characters are formed by typing their Romanized equivalents (Pinyin) on a standard QWERTY keyboard. Because Chinese is a homophonic language, many characters typically map to the same Pinyin representation, resulting in a palette of possibilities for the writer to choose among. (As an example, a few of the characters generated by entering the Pinyin “wu” follow, along with some of their contextually dependent meanings: 五[five], 午[noon], 务[business], 无[none], 舞[dance], 物[object], 吴[the surname, Wu], 乌[dark], and 伍[group].) In addition, possibilities may be displayed and selected before the Pinyin entry is complete. The displayed palettes may be context sensitive, ordering characters by the likelihood of fit with the immediately preceding text. A second method, Wubi, is based on the character’s structure, rather than its pronunciation. In Wubi, Roman letters map to character elements such that a character can be generated with four keystrokes at most. Wubi allows typing that is considerably faster than Pinyin. In essence, differences in the input methods used by individuals would need to be accounted for in attempting to measure writing processes. Fitting mixture distributions, for example, might be one approach to modeling fluency as a function of input method.

Several limitations of this study should be mentioned. First, we employed test-based writing in a single genre. That writing may differ in important ways from the writing required of educationally at-risk adults in post-secondary education or at work. Second, our process measures captured only a subset of the ones necessary for effective composition. In particular, our measures were more oriented toward basic processes like fluency than to such higher-order processes as idea generation. Finally, we were able to offer relatively little in terms of immediate implications of this research for instruction. More work needs to be done to identify the most promising uses of process data for enhancing teaching and learning.

References

Almond, R., Deane, P., Quinlan, T., Wagner, M., & Sydorenko, T. (2012). *A preliminary analysis of keystroke log data from a timed writing task*

(Research Report RR-12-23). Princeton, NJ: Educational Testing Service.

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

Bennett, R. E., Zhang, M., Deane, P., & van Rijn, P. W. (2020). How do proficient and less proficient students differ in their composition processes? *Educational Assessment*, 1–20.

Breslow, J. M. (2012, September). By the numbers: Dropping out of high school. *Frontline*.

Bridgeman, B., & Lewis, C. (1994). The relationship of essay and multiple-choice scores with grades in college courses. *Journal of Educational Measurement*, *31*, 37–50.

Camarata, S., & Woodcock, R. (2006). Sex differences in processing speed: Developmental effects in males and females. *Intelligence*, *34*, 231–252.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.

Connelly, V., Dockrell, J. E., Walter, K., & Critten, S. (2012). Predicting the quality of composition and written language bursts from oral language, spelling, and handwriting skills in children with and without specific language impairment. *Written Communication*, *29*, 278–302.

Deane, P. (2014). *Using writing process and product features to assess writing quality and explore how those features relate to other literacy tasks* (Research Report RR-14-03). Princeton, NJ: Educational Testing Service.

Deane, P., Roth, A., Litz, A., Goswami, V., Steck, F., Lewis, M., & Richter, T. (2018). *Behavioral differences between retyping, drafting, and editing: A writing process analysis* (Research Memorandum RM-18-06). Princeton, NJ: Educational Testing Service.

Deane, P., & Zhang, M. (2015). *Exploring the feasibility of using writing process features to assess text production skills* (Research Report RR-15-26). Princeton, NJ: Educational Testing Service.

for Education Sciences (IES), I. (2014). *US PIAAC prison study results: 2014* (Tech. Rep.). Retrieved from https://nces.ed.gov/surveys/piaac/results/prison_summary.aspx

Glaze, L. E., & Kaeble, D. (2014). *Correctional populations in the United States, 2013.* Washington, DC: US Department of Justice. Retrieved from https://www.aacu.org/sites/default/files/files/LEAP/2015_Survey_Report3.pdf

Guo, H., Deane, P. D., van Rijn, P. W., Zhang, M., & Bennett, R. E. (2018). Modeling basic writing processes from keystroke logs. *Journal of Educational Measurement*, *55*, 194–216.

Guo, H., Zhang, M., Deane, P., & Bennett, R. E. (2019). Writing process differences in subgroups reflected in keystroke logs. *Journal of Educational and Behavioral Statistics*, *44*, 571–596. https://doi.org/10.3102/1076998619856590.

Halpern, D. F. (2012). *Sex differences in cognitive abilities*. New York: Psychology Press.

Hart Research Associates. (2016). *Trends in learning outcomes assessment. Key findings from a survey among administrators at AAC&U member institutions* (Tech. Rep.). Retrieved from https://www.aacu.org/sites/default/files/files/LEAP/2015_Survey_Report3.pdf

Jewell, J., & Malecki, C. K. (2005). The utility of CBM written language indices: An investigation of production-dependent, production-independent, and accurate-production scores. *School Psychology Review*, *34*(1), 27–44.

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

Kellogg, R. T. (2001). Competition for working memory among writing processes. *The American Journal of Psychology*, *114*(2), 175.

Kuehl, R. O. (2000). *Design of experiments: statistical principles of research design and analysis* (3rd ed.). Pacific Grove, CA: Brookes/Cole.

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

Malecki, C. K., & Jewell, J. (2003). Developmental, gender, and practical considerations in scoring curriculum-based measurement writing probes. *Psychology in the Schools*, *40*(4), 379–390.

McCutchen, D. (1996). A capacity theory of writing: Working memory in composition. *Educational Psychology Review*, *8*(3), 299–325.

McCutchen, D. (2006). Cognitive factors in children's writing. In C. A. MacArthur, S. Graham, & F. Jill (Eds.), *Handbook of writing research.* New York: Guilford Press.

McCutchen, D. (2011). From novice to expert: Implications of language skills and writing-relevant knowledge for memory during the development of writing skill. *Journal of Writing Research*, *3*(1), 51–68. http://dx.doi.org/10.17239/jowr-2011.03.01.3.

McCutchen, D., Covill, A., Hoyne, S. H., & Mildes, K. (1994). Individual differences in writing: Implications of translating fluency. *Journal of Educational Psychology*, *86*(2), 256–266. https://doi.org/10.1037/0022-0663.86.2.256.

Norris, D., Oppler, S., Kuang, D., Day, R., & Adams, K. (2006). *The College Board SAT Writing validation study: An assessment of predictive and incremental validity* (Research Report 2006-2). New York: College Board. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.189.5460&rep=rep1&type=pdf

Pett, M. A., Lackey, N. R., & Sullivan, J. J. (2003). *Making sense of factor analysis: The use of factor analysis for instrument development in health care research*. Thousand Oaks, CA: Sage Publications Inc.

Reilly, D., Neumann, D. L., & Andrews, G. (2019). Gender differences in reading and writing achievement: Evidence from the National Assessment of Educational Progress (NAEP). *American Psychologist*, *74*(4), 445–458.

Révész, A., Michel, M., & Lee, M. (2019). Exploring second language writers'pausing and revision behaviors: A mixed-methods study. *Studies in Second Language Acquisition*, *41*, 605–631.

Scott, M. M., Zhang, S., & Koball, H. (2015). *Dropping out and clocking in: A portrait of teens who leave school early and work* (Low-Income Working Families Brief). Washington, DC: Urban Institute. Retrieved from https://www.urban.org/sites/default/files/publication/49216/2000189-Dropping-Out-and-Clocking-In.pdf

Strömqvist, S., Holmqvist, K., Johansson, V., Karlsson, H., & Wengelin, Å. (2006). What keystroke-logging can reveal about writing. In K. P. H. Sullivan & E. Lindgren (Eds.), *Computer keystroke logging and writing* (pp. 45–71). Oxford, UK: Elsevier.

Sullivan, K., & Lindgren, E. (2006). *Computer key-stroke logging and writing: Methods and applications*. Elsevier.

Tate, T. P., & Warschauer, M. (2019). Keypresses and mouse clicks: Analysis of the first national computer-based writing assessment. *Technology, Knowledge and Learning*, *24*, 523–543.

Wengelin, Å. (2006). Examining pauses in writing: Theory, methods and empirical data. In K. P. H. Sullivan & E. Lindgren (Eds.), *Computer keystroke logging and writing* (pp. 107–130). Oxford, UK: Elsevier.

Yan, C. M. W., McBride-Chang, C., Wagner, R. K., Zhang, J., Wong, A. M., & Shu, H. (2012). Writing quality in Chinese children: Speed and fluency matter. *Reading and Writing*, *25*, 1499–1521.

Zhang, M., Bennett, R. E., Deane, P., & 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., Deane, P., Feng, G., & Guo, H. (2019). *Investigating an approach to evaluating keyboarding fluency.* Paper presented at the 2019 Society for Text and Discourse (ST&D) Conference, New York, NY.

Zhang, M., Hao, J., Li, C., & Deane, P. (2018). Defining personalized writing burst measures of translation using keystroke logs. In K. E. Boyer & M. Yudelson (Eds.), *Proceedings of the 11th international conference on educational data mining* (pp. 549–552).

Zhang, M., Zhu, M., Deane, P., & Guo, H. (2019). Analyzing editing behaviors in writing using keystroke logs. In M. Wiberg, S. Culpepper, R. Janssen, J. Gonzalez, & D. Molenaar (Eds.), *Quantitative psychology: The 83rd annual meeting of the Psychometric Society.* New York: Springer.

Appendix: Writing-Process Indicators, Component Features, and Valences

Writing process indicator	Component feature valence	Component feature description
Between-word speed	+	Median rate of between-word whitespace keystrokes (in char/sec)
	−	Mean duration of between-word whitespace keystrokes (in logged ms)
	−	Median duration of the between-word whitespace keystroke intervals (in logged ms)
	−	Minimum duration of the between-word whitespace keystroke intervals (in logged ms)
	+	Mean rate of between-word append keystrokes, including white space and punctuation marks (in char/sec)
	+	Median rate of between-word append keystrokes, including white space and punctuation marks (in char/sec)
	−	Mean duration of the between-word append keystroke intervals (in logged ms)
	−	Median duration of the between-word append keystroke intervals (in logged ms)
	+	Median append-only burst length when burst is defined using 600ms as threshold for concluding pause length (in char)
Large-burst fluency	−	Relative probability of an error occurring during text generation (either a corrected typo or a final spelling error), on logit scale
	−	Relative probability of a word in the final text ever having been misspelled, on logit scale
	+	Mean append-only burst length when burst is defined using 8000ms as threshold for concluding pause length (in char)
	+	Median append-only burst length when burst is defined using 8000ms as threshold for concluding pause length (in char)
	+	Maximum append-only burst length when burst is defined using 8000ms as threshold for concluding pause length (in char)
	+	Maximum append-only burst length when burst is defined using 30000ms as threshold for concluding pause length (in char)
	−	Mean of all-action burst length where burst boundaries are defined as eight standard deviations above the median inter-key pause time (in char)
	+	Standard deviation of all append-only burst lengths when burst is defined using 8000ms as threshold for concluding pause length (in char)
	+	Mean of append-only burst length when burst is defined using 30000ms as threshold for concluding pause length (in char)
	+	Median of append-only burst length when burst is defined using 30000ms as threshold for concluding pause length (in char)
	+	Standard deviation of append-only burst length when burst is defined using 30000ms as threshold for concluding pause length (in char)
Productivity	−	Relative log odds of time spent on paste events compared to in-word events
	+	Total number of keystrokes
	+	Total number of between-sentence punctuation mark keystrokes
	+	Total number of in-word events
	+	Total number of append keystrokes
	+	Standard deviation of all-action burst length where burst boundaries are defined as eight standard deviations above the median inter-word pause time (in char)

Writing process indicator	Component feature valence	Component feature description
Productivity (con't)	+	Maximum all-action burst length where burst boundaries are defined as eight standard deviations above the median inter-word pause time (in char)
	+	Total number of word-initial keystroke events (i.e., an alpha-numeric character after a white space)
Deletion editing	+	Standard deviation of rate of long-distance jump-to-edit actions (in char/sec)
	+	Maximum rate of long-distance jump-to-edit actions (in char/sec)
	+	Total number of cut keystrokes
	+	Total time spent on cut events
	+	Relative time spent on cut events, on logit scale
	+	Relative log odds of time spent on cut events, compared to in-word time
	+	Standard deviation of cut keystrokes (in logged ms)
	+	Maximum cut keystroke latency (in logged ms)
	+	Maximum keystroke efficiency – number of characters/number of keystrokes per word
Jump editing	+	Relative log odds of time spent on jump-to-edit events within the same sentence, compared to in-word non-jump events
	+	Total time spent on jump-to-edit
	+	Total time spent on jump-to-edit events occurring within the same word
	+	Total time spent on long-distance jump-to-edit events
	+	Maximum duration of jump-to-edit keystrokes (in logged ms)
	+	Maximum duration of jump-to-edit keystrokes occurring within the same word (in logged ms)
	+	Mean jumped distance across jump-to-edit within the same word (in char)
	+	Standard deviation of jumped distances within the same sentence (in char)
	+	Maximum jumped distances within the same sentence (in char)
	+	Maximum duration of long-distance jump-to-edit keystrokes (in logged ms)
	+	Mean distance in long-distance jump-to-edit events (in char)
	+	Maximum duration of in-sentence punctuation keystrokes (in logged ms)
Between-sentence transition & backspace speed	+	Standard deviation of rate for between-sentence whitespace keystrokes (in char/sec)
	+	Maximum rate for between-sentence whitespace keystrokes (in char/sec)
	+	Maximum rate of inter-sentence interval keystrokes (in char/sec)
	+	Maximum rate of inter-sentence interval deletion keystrokes (in char/sec)
	+	Standard deviation of rate of continued backspace events (in char/sec)
	−	Minimum duration of continued backspace events (in logged ms)
Paragraphing	+	Total number of line-break keystrokes
	+	Relative time spent on line-break keystrokes, on logit scale
	+	Standard deviation of duration of line-break keystrokes (in logged ms)
	+	Max duration of line-break keystrokes (in logged ms)