Reading and Writing (2019) 32:2353–2374
https://doi.org/10.1007/s11145-019-09953-8

Understanding the keystroke log: the effect of writing task on keystroke features

Rianne Conijn[1] · **Jens Roeser**[2] · **Menno van Zaanen**[1]

Published online: 15 May 2019

Abstract

Keystroke logging is used to automatically record writers' unfolding typing process and to get insight into moments when they struggle composing text. However, it is not clear which and how features from the keystroke log map to higher-level cognitive processes, such as planning and revision. This study aims to investigate the sensitivity of frequently used keystroke features across tasks with different cognitive demands. Two keystroke datasets were analyzed: one consisting of a copy task and an email writing task, and one with a larger difference in cognitive demand: a copy task and an academic summary task. The differences across tasks were modeled using Bayesian linear mixed effects models. Posterior distributions were used to compare the strength and direction of the task effects across features and datasets. The results showed that the average of all interkeystroke intervals were found to be stable across tasks. Features related to the time between words and (sub)sentences only differed between the copy and the academic task. Lastly, keystroke features related to the number of words, revisions, and total time, differed across tasks in both datasets. To conclude, our results indicate that the latter features are related to cognitive load or task complexity. In addition, our research shows that keystroke features are sensitive to small differences in the writing tasks at hand.

Keywords Academic writing · Writing analytics · Keystroke logging · Bayesian linear mixed effects models · Typing characteristics

✉ Rianne Conijn
m.a.conijn@uvt.nl

Jens Roeser
jens.roeser@ntu.ac.uk

Menno van Zaanen
m.m.vanzaanen@uvt.nl

[1] Department of Cognitive Science and Artificial Intelligence, Tilburg University, Tilburg, The Netherlands

[2] Department of Psychology, Nottingham Trent University, Nottingham, UK

Introduction

Academic writing is an important skill in higher education and for the student's further professional career. Yet, several studies showed that students have difficulties with creating academic texts (e.g., Lea & Street, 1998; Mateos & Solé, 2009). Insight into students' writing processes can provide evidence on where and when students struggle (Likens, Allen, McNamara, 2017) and could be used to improve their writing ability (Deane, 2013). However, writing is a complex, highly recursive process, where different cognitive processes interact and can happen in any order. Flower and Hayes (1980)'s model distinguishes three main cognitive processes that interact: planning, translating, and reviewing. Given this complexity, it is difficult to provide automatic methods that allow insight into students' writing processes (Baaijen, Galbraith, & de Glopper, 2012; Leijten & Van Waes, 2013).

Writing processes have been measured in different ways, both during and after the writing process, using observations, video analysis, thinking-aloud methods, and retrospective interviews (e.g., Lei, 2008; Plakans, 2009; Xu & Ding, 2014). In the current study, we focus on the use of keystroke logging to measure writing processes during typing. With keystroke logging, the timing and type of every key press and key release are collected (Leijten & Van Waes, 2013). The analysis of these keystroke logs, keystroke analysis, is a promising area of research, because keystroke logs provide real-time, fine-grained information on writers' unfolding typing process during text composition. In addition, keystroke logs can be collected automatically, hence it is more scalable and less intrusive than traditional thinking-aloud methods and observation studies. Keystrokes have been used for a wide range of studies, including writer identification and authentication (Karnan, Akila, & Krishnaraj, 2011), prediction of performance in programming tasks (Thomas, Karahasanovic, & Kennedy, 2005), writing quality or essay scores (Zhang, Hao, Li, & Deane, 2016), writing fluency (Abdel Latif, 2009; Van Waes & Leijten, 2015), emotional states (Bixler & D'Mello, 2013; Salmeron-Majadas, Santos, & Boticario, 2014), deceptive writing (Banerjee, Feng, Kang, & Choi, 2014), task complexity (Grabowski, 2008), motor functionality (Van Waes, Leijten, Mariën, & Engelborghs, 2017), and linguistic features (Allen et al., 2016a). Moreover, several studies have shown that keystroke data can indeed be used for real-time information on the writing process (e.g., Tillema, van den Bergh, Rijlaarsdam, & Sanders, 2011; Baaijen et al., 2012; Van Waes, van Weijen, & Leijten, 2014).

These studies used a variety of keystroke features to inform their hypotheses. However, it is not yet clear how each of these features map onto underlying cognitive processes, which has been coined as the problem of alignment (Galbraith & Baaijen, 2019). Keystroke features may be multiply determined, and are sensitive to a variety of factors. In addition, keystroke features are not independent, and will, at least to some extent, overlap in the cognitive processes they are representing. Therefore, it is not always clear which features need to be selected for the question at hand. The selection of the correct keystroke features is crucial for the

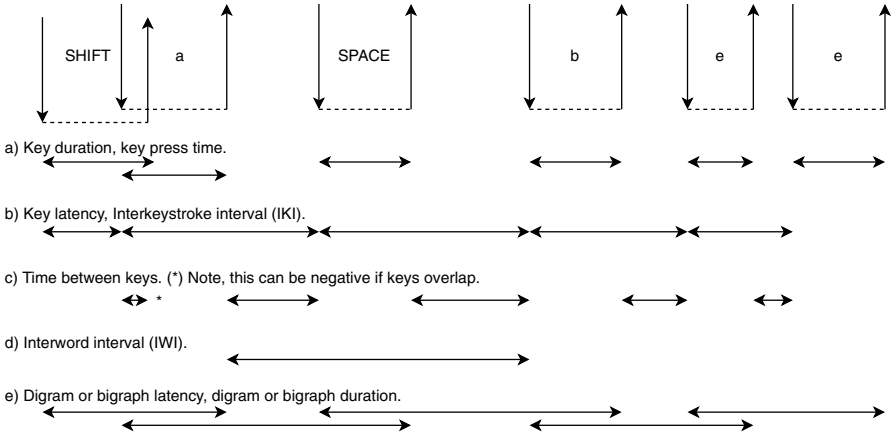


Fig. 1 Time-based features extracted from the keystroke log of typing: "A bee"

interpretation of the results and therefore, the derived conclusions. Hence, there is a need for a better understanding of sensitivity and independence of the keystroke features frequently used in the writing literature. In the current study we investigate the sensitivity of keystroke features across tasks.

Features extracted from keystroke logs

The features extracted from keystroke logs in previous work can be broadly organized into three categories: features related to the duration of the keystrokes, to content or revising behavior, and to written language bursts (e.g., Baaijen et al., 2012; Bixler & D'Mello, 2013). In the majority of studies the researchers extracted at least one or a few features related to duration, such as the duration between two consecutive key presses (e.g., Salmeron-Majadas et al., 2014) or the duration of one key press (e.g., Bixler & D'Mello, 2013; Allen et al., 2016b). The terminology of these time-based features is sometimes used interchangeably. For clarity, we provide an overview of the time-based features which are often extracted from keystrokes (see Fig. 1).

The specific duration features used depend on the hypothesis of the studies. For example, a literature review by Karnan et al. (2011) showed that the majority of writer identification and authentication studies focus on features such as key duration, keystroke latency, and digraph latency (see Fig. 1a, b, e). Sometimes the time between keys (see Fig. 1c) was included as well (e.g., Tappert, Villani, & Cha, 2009). For these features several summary statistics were computed, such as the mean and the standard deviation of the keystroke latencies. Also vectors representing keystroke features have been compared using Euclidean distance (Giot, El-Abed, & Rosenberger, 2009). All these measurements were computed for specific keys (e.g., "b"), and for specific combinations of keys (e.g., the bigram "be"). These combinations indicate how much time it takes to type a specific key, or sequence of keys.

In contrast, writing analytics studies focus on timing features in general, and not related to a specific key. The interkeystroke interval (IKI, see Fig. 1b) is the most commonly used feature. Multiple summary statistics are derived from IKIs, including the largest, smallest, mean, and median IKI. Note that these summary statistics are not merely used to describe the data, but as outcome variables representing the cognitive process of interest. Sometimes, these statistics are calculated per pause location; for example the IKI within or outside words (Grabowski, 2008), or the IKI within words, between words, between subsentences, and between sentences (Baaijen et al., 2012), or frequencies are extracted, such as the number of IKIs between 0.5–1.0 s, 1.0–1.5 s, 1.5–2.0 s, 2.0–3.0 s, and > 3.0 s (Allen et al., 2016a). In addition, other features related to time are extracted, such as initial pause time (e.g., Allen et al., 2016a), interword interval (e.g., Zhang et al., 2016), function or content word time (Banerjee et al., 2014), and total time (e.g., Bixler & D'Mello 2013).

Aside from duration features, content-related features were extracted, such as the number of keystrokes or verbosity (e.g., Allen et al., 2016b), the number of alphabetical keystrokes (Salmeron-Majadas et al., 2014), the number of words (e.g., Likens et al., 2017), the number of backspaces/deletes (e.g., Allen et al., 2016a), and efficiency, the number of characters in the final product per number of characters typed (Van Waes et al., 2014).

Lastly, some studies included features related to written language bursts. Writers compose sentence parts, identified by pauses longer than two seconds or by a grammatical discontinuity (Kaufer, Hayes, & Flower, 1986), also known as bursts. In keystroke analysis, written language bursts are often operationalized as sequences of text production (keystrokes) without an IKI longer than two seconds and without a revision and without an insertion away from the leading edge (Baaijen & Galbraith, 2018). Features related to written language bursts include the number of bursts or the number of words per burst after a pause, revision, or insertion (Baaijen et al., 2012).

Rationales for keystroke feature selection

The scientific rationales for selecting these keystroke features can be divided into data-driven and theory-driven approaches. On the one hand, studies using keystroke analysis for authentication and identification can be considered data-driven. These studies use multiple duration and content features to understand to what extent or with which accuracy the writer can be predicted, for example to build accurate automatic detection systems. Since including more features could lead to higher accuracy, often a combination is used of features that are known for their predictive power from previous studies, and 'new' features that are hypothesized to have predictive power (e.g., Karnan et al., 2011; Bixler & D'Mello, 2013). Since the main focus is on predictive accuracy, understanding the relation between the keystroke features and the cognitive processes is of limited interest in these studies. However, a large information gain of a keystroke feature on the prediction, found in multiple studies or contexts, could indicate a relationship worth investigating.

On the other hand, there are theory-driven approaches for keystroke feature selection. Several studies link the keystrokes to the three writing processes as defined by Flower and Hayes (1980): planning, translating, and reviewing processes. In addition, keystroke features have been related to cognitive load. Cognitive load reflects the notion that task performance is bound by the working memory capacity available for cognitive processing and the cognitive demands of a task (Sweller, 1988). If the cognitive demands of a task exceed the available working memory capacity, the writer might slow down, other (less demanding) strategies might be used, or more errors might be made (Just & Carpenter, 1992). In writing, high-level processes such planning and reviewing, are considered to have a high cognitive demand as they require high levels of attentional control (Alamargot, Dansac, Chesnet, & Fayol, 2007; Kellogg, 1996). In contrast, motor processes, such as typing, require less attention and hence have a lower demand Olive and Kellogg (2002).

Cognitive load, or planning and revising processes in general, are commonly related to duration features, such as the number of, length, and location of interkeystroke intervals or pauses (Wengelin, 2006; Van Waes et al., 2014). More pauses and longer pauses are related to a larger cognitive load (Wallot & Grabowski, 2013; Alves, Castro, De Sousa, & Strömqvist, 2007), that, for example, could indicate word and sentence planning or deliberation (Zhang et al., 2016; Roeser, Torrance, & Baguley, 2019). In contrast, shorter pauses are related to basic keyboard fluency or motor processes (Grabowski, 2008). Several studies also distinguish between different pause locations, such as between words or between (sub)sentences, or between and within words. Pauses before words are considered to reflect planning, retrieving, verifying, or editing processes, while pauses within words are considered to be related to typing skills (Grabowski, 2008; Baaijen et al., 2012). Pauses at sentence boundaries are considered to reflect global text planning and require more time, compared to pauses at word boundaries, which are considered to reflect lexical access Medimorec and Risko (2017). Features associated with content and revising content are frequently related to translation and revision processes. The number of words is often related to writing quality, where more words indicate a higher essay quality (e.g., Allen et al., 2016a). The number of deletions is argued to be related to revision processes (Van Waes et al., 2014), but also to lower-level aspects such as keyboard efficiency (Grabowski, 2008). Lastly, written language bursts are related to the execution process or Flower & Hayes's 1980 translation processes (Baaijen et al., 2012). Longer bursts and shorter pauses have been related to higher writing fluency and improved text quality (Alves & Limpo, 2015).

Thus, keystroke features are used to infer variations related to cognitive writing processes and cognitive demand required for these writing processes. However, it is unclear how exactly the features are related to these cognitive writing processes and how sensitive the features are to differences in cognitive demand Galbraith and Baaijen (2019). As different writing tasks are bound to reflect different cognitive demands, investigating differences in keystroke features across tasks could provide insight into the sensitivity of keystroke features to differences in cognitive demands.

Sensitivity of keystroke features across tasks

Previous studies showed that keystroke features differ between tasks. For example, features were shown to differ between a copy task (transcribing a text) and a—more demanding—email writing task (e.g., Tappert et al., 2009). However, these differences were not made explicit nor evaluated. Other studies did explicitly state the differences. Conijn and Van Zaanen (2017) found differences in the number of keystrokes, number of corrections, mean and standard deviation of interkeystroke interval within, before, and after word between an email writing and a copy task. Grabowski (2008) added a third task: copy from memory. Here, copying text was considered more difficult than copying text from memory, because the former task also included eye-hand coordination, needed for reading and reproducing the text. The most difficult task was email writing which involves planning and formulation in addition to motor-planning and execution. Results showed a larger efficiency (ratio between the number of characters in the final document and the number of keystrokes) for copy from memory, compared to copy from text and generation from memory. Typing speed, measured by interkeystroke interval between and within words, was found most stable across tasks. In the current study, we extend on this work by analyzing the differences in keystrokes across multiple tasks, which are assumed to differ in the required cognitive load and therefore, affecting keystroke features related to the cognitive processes involved.

Current study

We aim to investigate which, and how, keystroke features are affected by differences in cognitive load across writing tasks. This provides insight into the sensitivity of these features and which features are useful for analyzing cognitive writing processes. Finally, this could be used by educators or instructional designers to evaluate differences in cognitive demands imposed by their chosen learning designs.

Two datasets were collected, both containing keystroke data from two tasks: (1) Villani dataset, consisting of a copy task, where participants were asked to transcribe a given printed text, and an email writing task; and (2) Academic writing dataset, consisting of a copy task and an academic summary task. The copy task and the email writing task differ in terms of planning and revising processes. In a copy task, there will be no planning on a linguistic level, but only planning on a motor level (eye-hand coordination). In addition, revising will only take place for typos, but not for linguistic reasons. The copy task and the academic task differ even more in terms of planning and revising processes, compared to a copy task and an email writing task. This is because academic writing involves additional complexity, such as critical thinking, integrating sources, and utilizing a repertoire of linguistic practices appropriate for the task (Lea & Street, 1998).

Bayesian linear mixed models were used to determine the effect of these tasks on the keystroke features. Several keystroke features related to keystroke duration and deletions were extracted, because these have been related to cognitive load in general. We hypothesize that specifically features related to the time between words,

the time between sentences, and the amount of revision are sensitive to the tasks, as these are well-documented in the literature to be associated with cognitive writing processes (Wengelin, 2006; Van Waes et al., 2014). In contrast, we hypothesize that features related to keystroke duration within words are not sensitive to the tasks, because these have been associated with motor processes (Grabowski, 2008).

Method

Data were collected from two different datasets, both containing two different tasks: the Villani keystroke dataset, containing a copy task and an email writing task, and a dataset on academic writing recorded for the purpose of this research, containing a copy task and an academic summary task. The copy tasks differed to the extent that different texts were used to transcribe. However, both copy tasks did not require higher-level cognitive processes, such as planning on a linguistic level, involved in the other tasks.

Villani dataset

The open Villani keystroke dataset (Tappert et al., 2009; Monaco, Bakelman, Cha, & Tappert, 2012) is a keystroke dataset collected in an experimental setting. In the experiment, students and faculty could choose to conduct a copy task and/or an email writing task. The participants were allowed to type both forms of text multiple times. For the copy task, the participants were asked to copy a fable of 652 characters. In the email writing task, the participants were asked to write an arbitrary email of at least 650 characters. During the experiment, the key typed, time of key press, and time of key release were stored for every keystroke. In total, this resulted in more than one million keystrokes. The dataset consists of 142 participants, who wrote 359 copy texts and 1262 emails. The dataset and the collection of the dataset is explained in detail in Tappert et al. (2009).

For the current study, several data cleaning steps were taken. First, we only included data from participants who participated in both the copy task and the email writing task. This resulted in a dataset of 36 participants, who collectively wrote 338 copy texts and 416 emails. Second, inspections of the dataset showed some cases where a key was only released after a subsequent key was pressed, resulting in a negative time between keys. This for example happens when typing combination keys, such as SHIFT + {a–z} to capitalize a letter. Since we are interested in writing characteristics that differ across tasks, not in character-specific information such as capitalization, all times between keystrokes, words, subsentences, and sentences which were lower than 0, were coded as missing. Lastly, some participants typed only a few characters or clearly typed random sequences of characters, without spaces. Therefore, seven sessions were excluded where the number of keystrokes was smaller than 600 or the number of words was smaller than 50. This left us with a total of 747 sessions.

Academic writing dataset

The academic writing dataset was collected in an experimental setting; in an academic writing course for English second language learners. As part of the course, students were asked to complete two tasks: a copy task and an academic summary task. For the copy task, the students were asked to transcribe a fable of 850 characters. For the academic summary task, the students were asked to write a summary of 100–200 words based on a journal article. The journal article (Woong Yun & Park, 2011) described a 2×2 experimental design in the field of the students' major (Communication and Information Sciences). After reading the article, students were asked to write a summary within 30 min. During both tasks, keystrokes were collected using Inputlog (Leijten & Van Waes, 2013), from those students who provided informed consent. In total, 131 students participated in the study.

Similarly to the data cleaning of the Villani dataset, only data were included from participants who completed both the copy task and the summary task, resulting in data from 128 participants. In addition, all times between keystrokes, words, subsentences, and sentences which were lower than 0 were coded as missing. Lastly, since the summary task was considerably longer than the copy task, we only selected a subset of keystrokes of the summary task. Participants typed on average more than 900 characters in the copy task. Therefore, the first 900 characters were extracted from the summary task (session 1). If the participant wrote less than 900 characters, all characters were extracted. In addition, as most participants wrote more, the next 900 characters (901–1800) were also extracted from the summary task (session 2). For participants that wrote less than 1800 characters, all characters from character 901 were extracted, resulting in two subsets of keystrokes per participant. In addition, similarly to the Villani dataset, sessions were excluded where number of keystrokes were smaller than 600 or the number of words were smaller than 50. This resulted in a total of 128 copy task sessions and 115 (session 1) + 67 (session 2) = 182 summary task sessions.

Feature extraction

From both datasets, we extracted frequency-based and time-based features similar to those used in writing analytics literature. Five frequency-based features were extracted from the task as a whole, related to content and revision behavior: number of keystrokes, number of words, number of backspaces or deletes, efficiency (which is defined as the number of characters in the final document divided by the number of keystrokes), and the number of interkeystroke intervals (IKI) between 0.5 and 1.0 s. Although the number of IKIs larger than 1.0 s (IKIs between 1.0–1.5 s, 1.5–2.0 s, 2.0–3.0 s, and larger than 3.0 s) has been used as feature in previous writing studies (Bixler & D'Mello, 2013; Allen et al., 2016b), these were barely present in our dataset, and therefore not included in the present analysis.

Twenty time-based features were extracted. Seven of these were related to general keystroke durations, such as IKIs, the most commonly used feature in the literature,

including mean, standard deviation, median, largest, and smallest IKI (Fig. 1c), and the mean and standard deviation of the key press time (Fig. 1a). Additionally, time-based features were extracted, which were related to specific locations in the text, including the mean and standard deviation of IKI between words, IKI within word, the time between keys (Fig. 1c), the time between words or the interword interval (Fig. 1d), the time between sentences (indicated by periods, question marks, and exclamation marks), and the time between subsentences [indicated by commas, semicolons, and colons, as in Baaijen et al. (2012)]. Lastly, the total time of the task was computed. The time-based features showed large variation. To account for this positive skew, all time features (except for total time) were log transformed and all values above the 95th percentile were removed. Similar approaches were used in previous studies (e.g., Grabowski, 2008; Van Waes et al., 2017).

Analysis of differences in keystrokes between tasks

Bayesian linear mixed effects models (BLMMs; Gelman et al., 2014; Kruschke, 2014; McElreath, 2016) were used to determine the differences in keystroke features between tasks within each dataset and across the datasets. All keystroke features were used as dependent variables, and task (copy versus email writing and copy versus academic summary for the respective datasets) was added as a fixed effect. Participant ID was added as a random intercepts term accounting for variance in th keystroke featuers specific to individuals. In addition, the effect of task on the keystroke features might differ across less-experienced and more-experienced writers, as hypothesized by Grabowski (2008). This possibility was accounted for by adding by-participant slopes for task.

In the context of this study, a Bayesian approach was chosen for three reasons. First, BLMMs provide a reliable way of accounting for differences related to participant and task, with guaranteed convergence (Bates, Kliegl, Vasishth, & Baayen, 2015). Second, BLMMs make it possible to derive posterior probability distributions of the variables of interest (here, the task effect for each keystroke feature). Lastly, these posterior probability distributions can be used to compare the effect across various dependent variables within a dataset and, more importantly, across datasets.

For continuous models, linear models with log-normal distributions were used. For frequency data such as the number of words, distributions of the Poisson family were used. When discrete values were highly zero-inflated, e.g. included a large number of zero backspaces, negative binomials were used (Gelman & Hill, 2006; Gelman et al., 2014). Quasi-logit regressions were used for ratio data (efficiency), as these are bound between 0 and 1 (see e.g. Agresti, 2002; Barr, 2008; Donnelly and Verkuilen, 2017). In other words, the dependent variable was transformed from proportions to adjusted logits, and fitted as a continuous variable in linear regressions.[1]

[1] BLMMs were conducted in R using the R package "rstanarm" (Gabry & Goodrich, 2016). Weakly informative priors were used. The number of Markov chain Monte Carlo chains was set to 3 with 3,000 iterations per chain (1,500 warm-up). The Rubin-Gelman statistic (Gelman & Rubin, 1992), traceplots and leave-one-out cross-validation were used to determine model convergence (Vehtari, Gelman, & Gabry, 2015, Vehtari, Gelman, & Gabry, 2017).

The task effect (copy versus email/academic summary) on the keystroke features was evaluated in two ways. First, the most probable effect estimate $\hat{\beta}$ and its 95% credible interval were calculated to determine the size and direction of the effect. In contrast to confidence intervals, credible intervals indicate the range in which the true (unknown) parameter value (here, the task effect) lies with 95% probability (Kruschke, 2014; Nicenboim & Vasishth, 2016; Sorensen et al., 2016). If a credible interval includes zero, zero is a possible estimate of the effect of task on the outcome variable (the keystroke features).

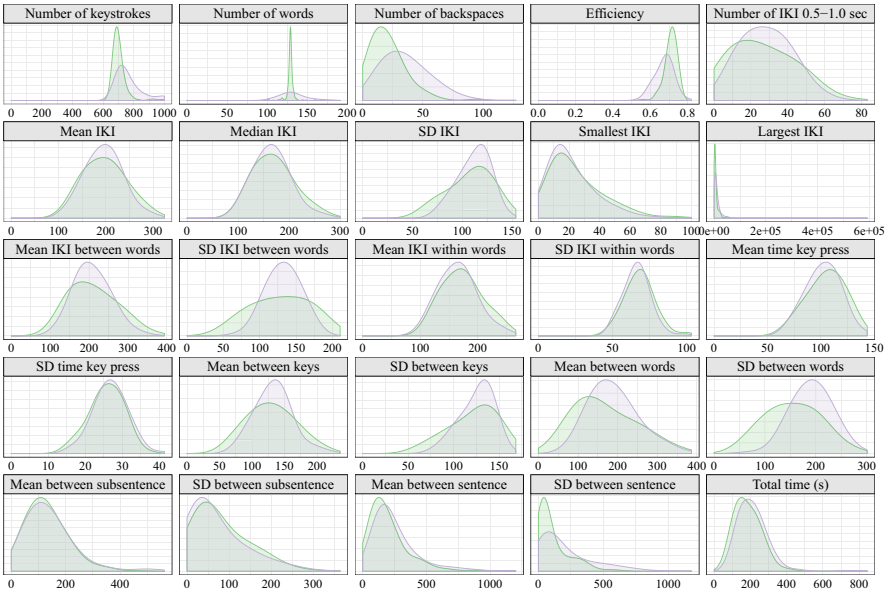
Second, the posterior probability distribution was used to calculate the standardized effect strength $\hat{\delta}$ which is defined as $\hat{\delta} = \frac{\hat{\beta}}{\hat{\sigma}}$, where $\hat{\beta}$ is the task effect estimate, and $\hat{\sigma}$ is the variance estimate for this effect. This effect strength allows us to compare the task effect across keystroke features within and across datasets, as it depends less on methodology or other experiment specific variables, such as language, text type, or participants, than the estimates (Wagenmakers, Lodewyckx, Kuriyal, & Grasman 2010).

Table 1 shows that task has an effect on several keystroke features. The direction of the task effect was largely similar across datasets. Specifically, most keystroke features showed a positive effect in both datasets, indicating larger values for the email writing task or the academic summary task, compared to the copy tasks. For example, participants paused 36 ms longer between words in the email writing task, compared to the copy task, and 208 ms longer between words in the academic summary task, compared to the copy task. Only efficiency and mean time key press were smaller for the email writing and the academic summary task. For the email writing task, efficiency was 41% lower and the mean key press time was 5 ms lower compared to the copy task. For the academic writing dataset, efficiency was 19% lower and mean key press time was 13 ms lower, compared to the copy task (Fig. 2). Thus, in the copy tasks, fewer keystrokes were needed per character in the final document and keys were pressed for a shorter period of time. In addition, the mean IKI within words and the smallest IKI were smaller for the email writing task, compared to the copy task, whereas the number of words and the mean IKI showed smaller values for the academic summary task.

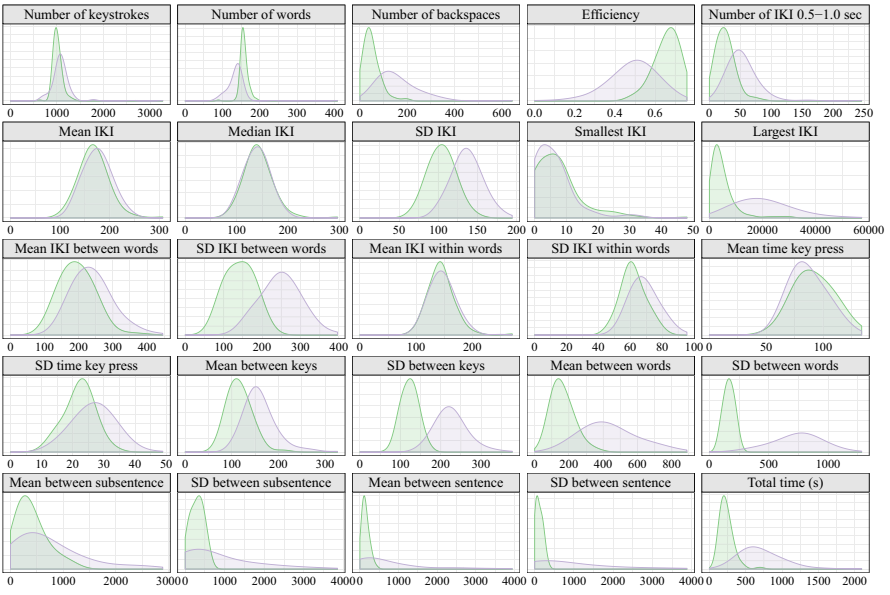
Interestingly, for the *SD* of time between subsentences, and mean IKI within and between words the most probable effect value ($\hat{\beta}$) changed in direction. Specifically, the *SD* of time between subsentences and the mean IKI within words and between words were lower for the copy task, compared to the email writing task in the Villani dataset. However, these were larger for the copy task, compared to the academic summary task in the academic writing dataset.

Effect strength of task

The posterior distributions of the effect strength are visualized in Fig. 3a for the Villani dataset, and in Fig. 3b for the academic writing dataset. The keystroke features were assigned to five groups that have been identified in previous studies: features related to the task in general, key presses, pauses in general (not location-specific),



(a) Villani dataset



(b) Academic writing dataset

Fig. 2 Distributions of the keystroke features per task (after trimming), for the Villani and the academic writing dataset. Note: All times are in ms [except total time (in s)]. For visualization purposes only, values larger than 4000 ms for the mean and *SD* time between sentences in the academic writing dataset were removed

Table 1 Task effects estimated from BLMMs on keystroke features

Keystroke feature	Villani dataset			Academic writing dataset		
	Lower	$\hat{\beta}$	Upper	Lower	$\hat{\beta}$	Upper
Number of keystrokes	1.01	1.04	1.08	1.38	1.45	1.51
Number of words	− 1.06	− 0.99	− 0.93	− 0.78	− 0.73	− 0.68
Number of backspaces	1.02	1.12	1.22	1.27	1.35	1.42
Efficiency	− 0.44	− 0.41	− 0.39	− 0.22	− 0.19	− 0.16
Number of IKI 0.5–1.0 sec	1.11	1.51	1.97	4.69	8.46	12.88
Mean IKI	− 11.77	− 2.70	5.61	− 16.54	− 10.74	− 5.21
Median IKI	1.02	1.12	1.22	1.27	1.35	1.42
SD IKI	0.06	0.11	0.16	0.52	0.60	0.69
Smallest IKI	− 7.65	− 4.26	− 0.80	− 2.45	− 1.15	0.25
Largest IKI (s)	8.81	17.14	27.27	133.93	189.57	247.65
Mean IKI between words	− 12.83	− 0.19	12.43	20.31	27.13	34.08
SD IKI between words	0.05	0.11	0.18	0.63	0.71	0.80
Mean IKI within words	− 18.54	− 11.69	− 4.52	− 0.63	2.71	6.18
SD IKI within word	− 0.00	0.04	0.09	0.02	0.06	0.09
Mean time key press	− 7.54	− 4.55	− 1.49	− 14.86	− 12.64	− 10.52
SD time key press	0.02	0.05	0.07	0.26	0.31	0.36
Mean between keys	1.87	10.12	19.04	19.72	25.38	30.66
SD between keys	0.01	0.11	0.20	1.26	1.48	1.70
Mean between words	19.16	36.31	53.54	172.40	208.28	246.29
SD between words	0.14	0.30	0.47	2.96	3.37	3.84
Mean between subsentence	− 23.14	6.66	44.30	121.39	284.03	460.96
SD between subsentence	− 0.39	− 0.12	0.16	− 0.32	0.44	1.19
Mean between sentence	16.35	74.10	138.60	470.47	1056.81	1778.19
SD between sentence	0.10	0.39	0.70	10.92	18.56	27.39
Total time (s)	51.56	80.05	108.14	2692.67	3295.20	3853.06

All values are shown in their original units. All times are in ms (except from total time and largest IKI). Positive values indicate larger values for the email writing or academic summary task and negative values indicate larger values for the copy tasks. $\hat{\beta}$ is the most probable estimate for the difference between tasks. Lower and upper specifies the 95% credible interval around the estimate $\hat{\beta}$

pauses within words, and pauses between words (Grabowski, 2008; Wallot & Grabowski, 2013).

For the Villani dataset, firstly, the task effect was largest for features related to the task as a whole, with the largest positive effect on the number of backspaces and total time, and the largest negative effect on the efficiency. Thus, the email writing task consisted of more backspaces, a lower ratio of characters typed to the characters in the final product, and took longer to type, compared to the copy task. Second, features related to pausing, such as mean and median IKI, showed low effect strengths, while features related to the variance in the general pauses, such as the largest IKI and *SD* IKI, showed large effect strengths. Third, the effect

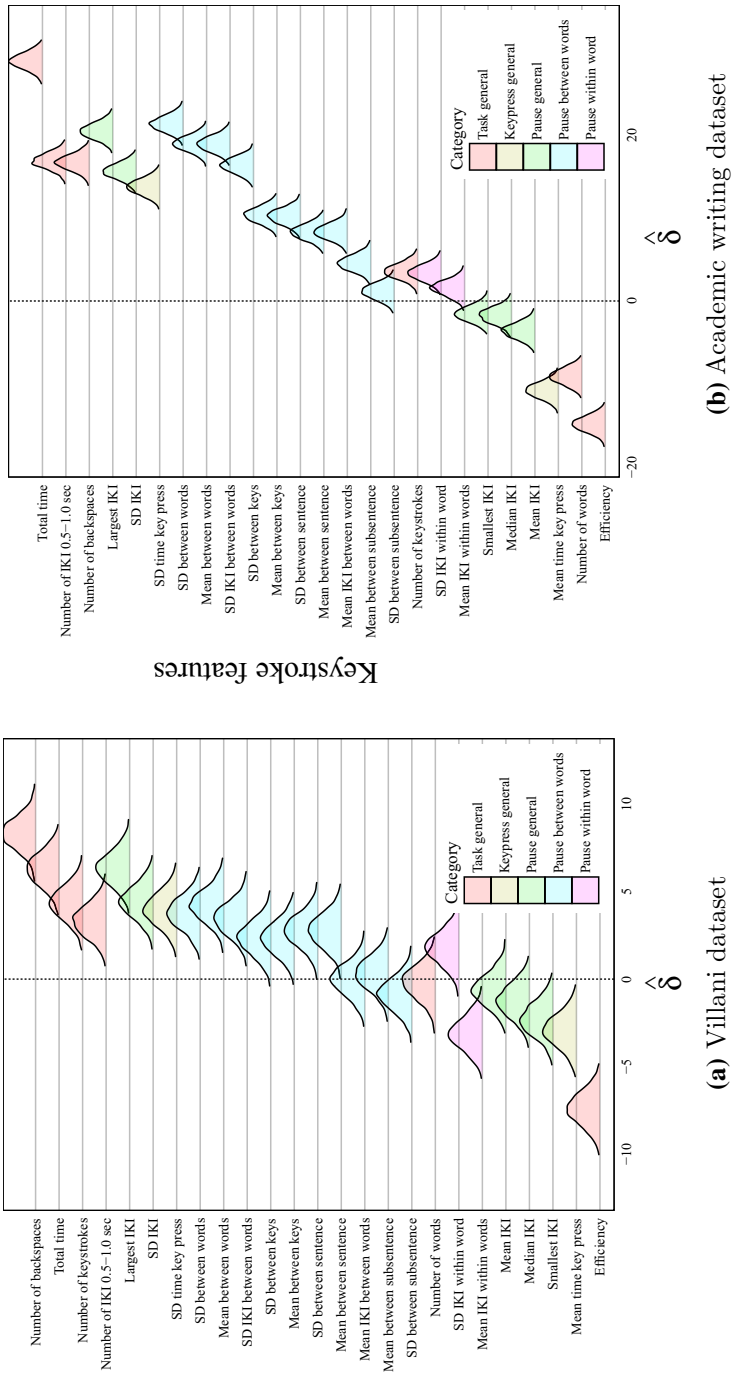


Fig. 3 Task effect strength $\hat{\delta}$ (academic summary/email writing—copy task) per keystroke feature for both datasets. Distributions are grouped by category and effect strength

on location-specific pause features was relatively low, except from the mean and the *SD* of time and the *SD* IKI between words. Fourth, the mean IKI within words showed a small negative effect, while the effect on the variance between words was positive. This indicates that participants typed faster within words, but with a larger variance, in the email task compared to the copy task. Lastly, the mean key press time showed a negative effect, while the *SD* key press time showed a positive effect.

For the academic writing dataset, firstly, again the task effect was largest for features related to the task as a whole. Total time showed a large positive effect, and efficiency showed the largest negative effect. Thus, the academic writing task took longer, and had a lower ratio of characters typed to the characters in the final product, compared to the copy task. Second, for the features related to pauses in general, a large effect was found for the *SD* and largest IKI, but a small effect for the mean and median IKI. Third, the features related to location-specific pauses between words and between sentences showed relatively large effect strengths, especially for the mean and *SD* of time and the *SD* IKI between words. The effect strengths of the mean and *SD* of time between (sub)sentences were considerably smaller. Fourth, the pauses within words showed little effect. Lastly, the *SD* key press time showed a positive effect, and a slight negative effect for the mean key press time.

Conceptually related features showed similar patterns within the dataset. For example, for both mean and median IKI, the difference between tasks could be both positive and negative, rendering a small effect strength. In addition, the *SD* of time between words and the *SD* IKI between words showed a positive effect within both datasets, with similar effect sizes. However, the mean IKI between words and mean time between words did not show similar patterns in the Villani dataset: the effect of task on the mean time between words was positive with a relatively large effect strength, while the effect of task on the mean IKI between words had a small effect strength, where the direction could not be determined.

Differences in effect strength between datasets

The five groups of keystrokes showed similar patterns in effect strengths between the two datasets. However, the actual effect strength of task differed across datasets: in the academic writing dataset, the effect strengths were larger for almost all keystroke features, compared to the Villani dataset. These differences are shown in Fig. 4. Comparable effects strengths across datasets (reliable effects) were found for the number of keystrokes, median IKI, smallest IKI, and *SD* IKI between words. Total time showed the largest difference in effect strength across datasets, indicating a task specific effect (email writing/academic writing), rather than a task general effect on total time. In addition, especially the variances in pause times (*SD* time between words, *SD* time between keys, *SD* IKI between words, *SD* time between sentences, *SD* IKI, and *SD* time key press) showed large differences in effect. Moreover, task general effects such as efficiency, number of backspaces and number of words showed large differences in effect.

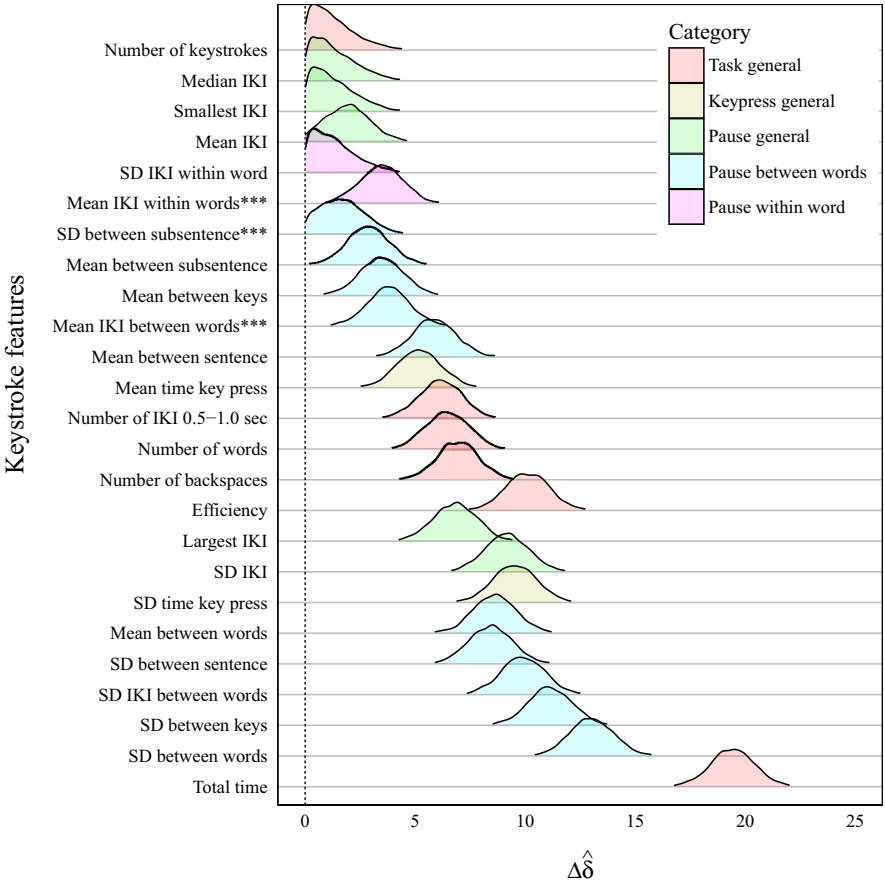


Fig. 4 Absolute difference Δ of the effect strength $\hat{\delta}$ contrasting both data sets. Differences are shown by keystroke feature. Values close to zero indicate similar effects in both datasets. Larger values indicate effects that are different for and thus, specific to the email/academic writing task. Distributions are grouped by category and effect strength. *** Keystroke features for which the direction of effect differed between datasets

Discussion

In this study we aimed to investigate which, and how, keystroke features are affected by differences in cognitive demand across writing tasks. To achieve this we extracted various keystroke features which are related to pause durations (general and location-based), and content and revising behavior. The keystrokes were compared across two different datasets, both containing a copy task and one containing an email writing task and the other an academic summary task. Bayesian linear mixed effects models were applied to determine the strength and direction of the effects of task between the different keystroke features within and across datasets.

Some keystrokes showed an effect of task in both datasets, some in only one dataset, and some did not show an effect in either dataset.

First, several keystrokes features differed between the tasks in both datasets. It was hypothesized that features related to the time between words and sentences, and the amount of revisions would differ across tasks, because these are frequently associated with cognitive processes, such as planning and revising (Wengelin, 2006; Van Waes et al., 2014). This was confirmed in both datasets. In particular, features related to the task as a whole, such as the number of keystrokes, the number of backspaces, efficiency, largest IKI, and total time, were different between the two tasks in both datasets. In addition, the mean time between words differed between writing tasks. These findings reproduced across not just the present datasets, but were also reported by other studies (Grabowski, 2008; Conijn & Van Zaanen, 2017). These features seem to be strongly influenced by the writing task, are not specific to datasets and, therefore, must be sensitive to task characteristics, such as cognitive demand. This allows the use of these features for task classification, at least for the tasks reported in those studies.

Second, some keystroke features related to time between words and sentences only showed differences in effect in the academic dataset, but not in the Villani dataset. The mean of the IKI between words, time between keys, time between subsentences, and the standard deviation of the IKI between words, time between keys, and time between subsentences, and the number of words only differed between the academic summary task and the copy task, but not between the copy task and the email writing task. One possible explanation for this is that these features are only affected by the task, if the difference in the cognitive demands are larger. In the present datasets, the academic summary task could be considered more complex compared to the email writing task, because it involves additional complexity, such synthesizing, integrating sources, and utilizing a repertoire of linguistic practices appropriate for the task (Lea & Street, 1998). Therefore, these features might be less sensitive to small differences in complexity or cognitive demand.

Third, it was hypothesized that keystroke duration within words would not be sensitive to task because these are associated with motor processes or individual typing skills (Grabowski, 2008). Indeed, it was shown that the mean and standard deviation of the interkeystroke intervals within words did not differ between tasks in the academic writing dataset. This could indicate that the cognitive writing processes during word production, beyond motor processes and typo revisions, are limited, or that cognitive writing processes within words are reflected similarly in the interkeystroke intervals within words in both tasks.

In addition, we found that conceptually related keystroke features, such as mean and median interkeystroke intervals, had similar effects within the dataset. Interestingly, the effect of task on the mean time between words and the mean interkeystroke interval between words differed in the Villani dataset: the mean time between words showed a positive effect of task with a relatively large effect strength, while the mean interkeystroke interval between words showed that the effect could be both positive and negative, with a really small effect strength. A possible explanation lies in the different measurements of these features. The mean time between words is the whole pause between words, while the interkeystroke interval between words only

measures the interkeystroke interval of the last letter of the word and the 'space' key pressed (see Fig. 1). This would suggest that the feature time between words more easily picks up on the differences in task, compared to the somewhat lower-level feature interkeystroke intervals between words.

Extending on earlier research, we not only showed which keystroke features differed across tasks, but also compared the strength and the direction of the effects within and across dataset. For the Villani dataset, the effect of task was largest for the number of backspaces, the largest interkeystroke interval, total time, and efficiency. Thus, in the email writing task more backspaces were used, the largest interkeystroke interval was longer, the total time spent was longer, and the efficiency was lower, compared to the copy task. In the academic writing dataset, the largest effects were found for total time, *SD* between words, the largest interkeystroke interval, and efficiency. This indicates that students spent more time, had more variance in time between words, longer largest interkeystroke intervals, and lower efficiency in the academic summary task, compared to the copy task.

When comparing the effects across datasets, it was found that total time, *SD* of time between keys and between words are the keystroke features which differ most in the effect of task across the two datasets. Since the two datasets were assumed to vary in terms of the complexity of the writing, as opposed to copy task, this might indicate the usefulness of these features for determining task complexity or cognitive demand. The *SD* of time between subsentences and mean interkeystroke interval between and within words even differed in direction of the effect across the datasets. The change in direction of the effect of task across datasets might indicate that these features are more related to the specific dataset rather than to the effect of task. For example, the language or style could be more complex in the Villani copy text compared to the email text, while the language or style in the academic dataset copy task could be less complex compared to the academic writing task.

This study is limited in three ways. First, we compared two tasks in two different datasets. We argued that some of the differences in keystroke features might be due to the task complexity or cognitive demand, which differed across tasks. However, this might also be caused by other task characteristics, which we did not measure. The copy tasks were non-identical. Nevertheless, because both copy tasks did not require higher-level cognitive processes, such as linguistic planning, the differences can still be explained by the task complexity or cognitive demand. In addition, the differences might be due to other task characteristics, such as required style. However, for the purpose of this paper we were not interested why the keystroke features differed, but merely which and how.

Second, we did not explicitly measure the complexity or the cognitive load demand of the task. Thus, we cannot specifically state the exact relation between cognitive load and the keystroke features. For example, we do not know whether the relation between time between words and cognitive load will be linear. Although beyond the scope of the current study, it would be interesting for future work to further investigate the influence of cognitive load on the keystroke features. This could be done by comparing the keystroke features of multiple tasks of which the cognitive load or complexity is known, for example, by using a secondary task or questionnaire (e.g., Paas et al. 2003). This information might also

be used to identify when a task is too complex or requires too much cognitive load.

Third, the differences in the keystroke features might be caused by other factors that we did not test. Keystrokes are found to be sensitive to other factors, such as handedness, keyboard type (Gunetti & Picardi, 2005; Tappert et al., 2009), typing and writing experience and abilities, environmental conditions (Gunetti & Picardi, 2005), and cognitive impairments (e.g., Van Waes et al., 2017). We do know that the participant samples differed between the datasets (students versus students and faculty), which might indicate differences in writing experience. Yet, participant specific variation was statistically accounted for, so the differences across the dataset could not be explained by individual differences in the samples. However, the same approach used in the current study could be used in future work to identify the influence of these other factors on the keystroke features. In this way, we could identify which and how keystroke features are sensitive to individual differences and experimental factors. This could indicate which factors need to be controlled for when analyzing specific keystroke features. For example, when handedness does not appear to influence the number of backspaces, handedness does not need to be controlled for when analyzing the effect of the number of backspaces between writers on the dependent variable of interest.

Although previous work has hypothesized that some of these features are related to cognitive demand, in this study we specifically showed which and how these features differed with different cognitive demands across tasks. These findings provide insight into which features are of interest when we are looking for evidence of cognitive writing processes, such as planning, translating, and reviewing processes (Flower & Hayes, 1981), in the keystroke log. In addition, the sensitivity of the keystroke features across tasks shows that caution should be taken when generalizing the effect of these features across tasks, because these features might differ merely as a result of the task, rather than as a result of the variable of interest, for example, writing quality, which has frequently been predicted in writing research (e.g., Allen et al., 2016a; Zhang et al., 2016; Likens et al., 2017).

Next to these theoretical implications, the findings of the current study have implications for educational practice. This study showed which keystroke features differ across tasks with different cognitive demands, and hence might be used as to determine differences in cognitive load between tasks. Educators and instructional designers could use these insights to identify differences in cognitive demands imposed by their chosen learning designs. This would allow them to automatically evaluate whether their chosen writing tasks are producing the expected learning processes and outcomes (Kennedy & Judd, 2007; Lockyer, Heathcote, & Dawson, 2013). In addition, as keystrokes are measured during the writing process, differences in cognitive load might be determined during a single task. This could be used, for example, to determine cognitive load during different writing processes, such as planning, translating, and reviewing (cf. Alves, Castro, & Olive, 2008). These insights, could be used by teachers to when of with which writing processes a student could use support to improve their writing process (Santangelo, Harris, & Graham, 2016).

Conclusion

To conclude, this study provided insight into how keystroke-based features differ across writing tasks with different cognitive demands. Features related to interkeystroke intervals in general, or interkeystroke intervals within words did not differ across task. Features related to the time between words or sentences, such as *SD* interkeystroke between sentences, or mean interkeystroke interval between words, only differed between tasks with larger differences in cognitive demands. Lastly, features related to task as a whole, such as the number of words typed, amount of revision, and total time, as well as the time between words were found to differ across all tasks. This indicates that especially these latter features are related to cognitive load or task complexity, and hence would be of interest for analyzing cognitive writing processes. In addition, this study showed that it is important to be mindful when deriving conclusions from individual keystroke features, because they are already sensitive to small differences in writing tasks. To conclude, this study provides us with a better understanding of the keystroke features frequently used in the writing literature.

References

Abdel Latif, M. M. (2009). Toward a new process-based indicator for measuring writing fluency: Evidence from L2 writers' think-aloud protocols. *Canadian Modern Language Review*, *65*(4), 531–558. https://doi.org/10.3138/cmlr.65.4.531.

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). Hoboken, NJ: Wiley. https://doi.org/10.1007/978-3-642-04898-2_161.

Alamargot, D., Dansac, C., Chesnet, D., & Fayol, M. (2007). Parallel processing before and after pauses: A combined analysis of graphomotor and eye movements during procedural text production. In M. Torrance, L. Van Waes, & D. Galbraith (Eds.), *Studies in writing* (pp. 13–29). Bingley: Emerald Group Publishing.

Allen, L. K., Jacovina, M. E., Dascalu, M., Roscoe, R. D., Kent, K., Likens, A. D. & McNamara, D. S. (2016a). ENTER ing the Time Series SPACE: Uncovering the Writing Process through Keystroke Analyses. In *Proceedings of the 9th international conference on educational data mining (EDM)* (pp. 22–29). https://eric.ed.gov/?id=ED592674.

Allen, L. K., Mills, C., Jacovina, M. E., Crossley, S., D'Mello, S. & McNamara, D. S. (2016b). Investigating boredom and engagement during writing using multiple sources of information: the essay, the writer, and keystrokes. In *Proceedings of the 6th international conference on learning analytics & Knowledge* (pp. 114–123). https://doi.org/10.1145/2883851.2883939.

Alves, R. A., Castro, S. L. , De Sousa, L., & Strömqvist, S. (2007). Influence of typing skill on pause-execution cycles in written composition. In: Writing and cognition: Research and applications (pp. 55–65). Brill Nijhoff: Brill.

Alves, R. A., Castro, S. L., & Olive, T. (2008). Execution and pauses in writing narratives: Processing time, cognitive effort and typing skill. *International Journal of Psychology*, *43*(6), 969–979. https://doi.org/10.1080/00207590701398951.

Alves, R. A., & Limpo, T. (2015). Progress in written language bursts, pauses, transcription, and written composition across schooling. *Scientific Studies of Reading*, *19*(5), 374–391. https://doi.org/10.1080/10888438.2015.1059838.

Baaijen, V. M. & Galbraith, D. (2018). Discovery through writing: Relationships with writing processes and text quality. *Cognition and Instruction,* 1–25. https://doi.org/10.1080/07370008.2018.1456431.

Baaijen, V. M., Galbraith, D., & de Glopper, K. (2012). Keystroke analysis: Reflections on procedures and measures. *Written Communication*, *29*(3), 246–277. https://doi.org/10.1177/0741088312451108.

Banerjee, R., Feng, S., Kang, J. S., & Choi, Y. (2014). Keystroke patterns as prosody in digital writings: A case study with deceptive reviews and essays. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 1469–1473). https://doi.org/10.3115/v1/D14-1155.

Barr, D. J. (2008). Analyzing 'visual world' eyetracking data using multilevel logistic regression. *Journal of Memory and Language*, *59*(4), 457–474. https://doi.org/10.1016/j.jml.2007.09.002.

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. *arXiv preprint*. arXiv:org/abs/1506.04967.

Bixler, R. & D'Mello, S. (2013). Detecting boredom and engagement during writing with keystroke analysis, task appraisals, and stable traits. In: Proceedings of the 2013 international conference on intelligent user interfaces (IUI) (pp. 225–234). New York, NY, USAACM. https://doi.org/10.1145/2449396.2449426.

Conijn, R., & Van Zaanen, M., (2017). Identifying writing tasks using sequences of keystrokes. In: Benelearn,. (2017). *Proceedings of the 26th Benelux conference on machine learning benelearn* (pp. 28–35). Eindhoven: The Netherlands.

Deane, P. (2013). On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing*, *18*(1), 7–24. https://doi.org/10.1016/j.asw.2012.10.002.

Donnelly, S., & Verkuilen, J. (2017). Empirical logit analysis is not logistic regression. *Journal of Memory and Language*, *94*, 28–42. https://doi.org/10.1016/j.jml.2016.10.005.

Flower, L., & Hayes, J. R. (1980). The cognition of discovery: Defining a rhetorical problem. *College Composition and Communication*, *31*(1), 21–32. https://doi.org/10.2307/356630.

Flower, L., & Hayes, J. R. (1981). A cognitive process theory of writing. *College Composition and Communication*, *32*(4), 365–387. https://doi.org/10.2307/356600.

Gabry, J. & Goodrich, B. (2016). *RSTANARM: Bayesian applied regression modeling via Stan (Computer software manual)*. https://CRAN.R-project.org/package=rstanarm (R package version 2.13.1)

Galbraith, D. & Baaijen, V. M. (2019). Aligning keystrokes with cognitive processes in writing. In *Observing writing* (pp. 306–325). Brill. https://doi.org/10.1163/9789004392526_015.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (3rd ed.). New York: Chapman and Hall/CRC.

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge: Cambridge University Press.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*, 457–472. https://doi.org/10.1214/ss/1177011136.

Giot, R., El-Abed, M., & Rosenberger, C. (2009). Keystroke dynamics authentication for collaborative systems. In *International symposium on collaborative technologies and systems* (pp. 172–179). https://doi.org/10.1109/CTS.2009.5067478.

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

Gunetti, D., & Picardi, C. (2005). Keystroke analysis of free text. *ACM Transactions on Information and System Security (TISSEC)*, *8*(3), 312–347. https://doi.org/10.1145/1085126.1085129.

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*, *99*(1), 122–149. https://doi.org/10.1037/0033-295X.99.1.122.

Karnan, M., Akila, M., & Krishnaraj, N. (2011). Biometric personal authentication using keystroke dynamics: A review. *Applied Soft Computing*, *11*(2), 1565–1573. https://doi.org/10.1016/j.asoc.2010.08.003.

Kaufer, D. S., Hayes, J. R., & Flower, L. (1986). Composing written sentences. *Research in the Teaching of English*, pp. 121–140. https://www.jstor.org/stable/40171073.

Kellogg, R. T. (1996). A model of working memory in writing. In *The Science of Writing. Theories, Methods, Individual Differences, and Applications* (pp. 57–71).

Kennedy, G. E., & Judd, T. S. (2007). Expectations and reality: Evaluating patterns of learning behaviour using audit trails. *Computers & Education*, *49*(3), 840–855. https://doi.org/10.1016/j.compedu.2005.11.023.

Kruschke, J. (2014). *Doing bayesian data analysis: A tutorial with R, JAGS, and STAN*. Cambridge: Academic Press.

Lea, M. R., & Street, B. V. (1998). Student writing in higher education: An academic literacies approach. *Studies in Higher Education*, *23*(2), 157–172. https://doi.org/10.1080/03075079812331380364.

Lei, X. (2008). Exploring a sociocultural approach to writing strategy research: Mediated actions in writing activities. *Journal of Second Language Writing*, *17*(4), 217–236. https://doi.org/10.1016/j.jslw.2008.04.001.

Leijten, M., & Van Waes, L. (2013). Keystroke logging in writing research: Using inputlog to analyze and visualize writing processes. *Written Communication*, *30*(3), 358–392. https://doi.org/10.1177/0741088313491692.

Likens, A. D., Allen, L. K. & McNamara, D. S. (2017). Keystroke Dynamics Predict Essay Quality. In *Proceedings of the 39th annual meeting of the cognitive science society (CogSci 2017)* (pp. 2573–2578). London: UK.

Lockyer, L., Heathcote, E., & Dawson, S. (2013). Informing pedagogical action: Aligning learning analytics with learning design. *American Behavioral Scientist*, *57*, 1439–1459. https://doi.org/10.1177/0002764213479367.

Mateos, M., & Solé, I. (2009). Synthesising information from various texts: A study of procedures and products at different educational levels. *European Journal of Psychology of Education*, *24*(4), 435–451. https://doi.org/10.1007/BF03178760.

McElreath, R. (2016). *Statistical rethinking: A bayesian course with examples in R and Stan*. Boca Raton: CRC Press.

Medimorec, S., & Risko, E. F. (2017). Pauses in written composition: On the importance of where writers pause. *Reading and Writing*, *30*(6), 1267–1285. https://doi.org/10.1007/s11145-017-9723-7.

Monaco, J. V., Bakelman, N., Cha, S.- H., & Tappert, C. C. (2012). Developing a keystroke biometric system for continual authentication of computer users. In *Intelligence and Security Informatics Conference (EISIC), 2012 European* (pp. 210–216). https://doi.org/10.1109/EISIC.2012.58.

Nicenboim, B., & Vasishth, S. (2016). Statistical methods for linguistic research: Foundational Ideas-Part II. *Language and Linguistics Compass*, *10*(11), 591–613. https://doi.org/10.1111/lnc3.12207.

Olive, T., & Kellogg, R. T. (2002). Concurrent activation of high-and low-level production processes in written composition. *Memory & Cognition*, *30*(4), 594–600. https://doi.org/10.3758/BF03194960.

Paas, F., Tuovinen, J. E., Tabbers, H., & Van Gerven, P. W. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist*, *38*(1), 63–71. https://doi.org/10.1207/S15326985EP3801_8.

Plakans, L. (2009). Discourse synthesis in integrated second language writing assessment. *Language Testing*, *26*(4), 561–587. https://doi.org/10.1177/0265532209340192.

Roeser, J., Torrance, M., & Baguley, T. (2019). Advance planning in written and spoken sentence production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. https://doi.org/10.1037/xlm0000685.

Salmeron-Majadas, S., Santos, O. C., & Boticario, J. G. (2014). An evaluation of mouse and keyboard interaction indicators towards non-intrusive and low cost affective modeling in an educational context. *Procedia Computer Science*, *35*, 691–700. https://doi.org/10.1016/j.procs.2014.08.151.

Santangelo, T., Harris, K., & Graham, S. (2016). Self-regulation and writing. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 174–193). New York: The Guilford Press.

Sorensen, T., Hohenstein, S., & Vasishth, S. (2016). Bayesian linear mixed models using stan: A tutorial for psychologists, linguists, and cognitive scientists. *The Quantitative Methods for Psychology*, *12*(3), 175–200. https://doi.org/10.20982/tqmp.12.3.p175.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, *12*(2), 257–285. https://doi.org/10.1207/s15516709cog1202_4.

Tappert, C. C., Villani, M., & Cha, S.- H. (2009). Keystroke biometric identification and authentication on long-text input. *Behavioral biometrics for human identification: Intelligent applications*. https://doi.org/10.4018/978-1-60566-725-6.ch016.

Thomas, R. C., Karahasanovic, A., & Kennedy, G. E. (2005). An investigation into keystroke latency metrics as an indicator of programming performance. In *Proceedings of the 7th Australasian conference on Computing education-Volume 42* (pp. 127–134). Australian Computer Society, Inc. http://dl.acm.org/citation.cfm?id=1082440.

Tillema, M., van den Bergh, H., Rijlaarsdam, G., & Sanders, T. (2011). Relating self reports of writing behaviour and online task execution using a temporal model. *Metacognition and Learning*, *6*(3), 229–253. https://doi.org/10.1007/s11409-011-9072-x.

Van Waes, L., & Leijten, M. (2015). Fluency in writing: A multidimensional perspective on writing fluency applied to l1 and l2. *Computers and Composition*, *38*, 79–95. https://doi.org/10.1016/j.compcom.2015.09.012.

Van Waes, L., Leijten, M., Mariën, P., & Engelborghs, S. (2017). Typing competencies in Alzheimer's disease: An exploration of copy tasks. *Computers in Human Behavior*, *73*, 311–319. https://doi.org/10.1016/j.chb.2017.03.050.

Van Waes, L., van Weijen, D., & Leijten, M. (2014). Learning to write in an online writing center: The effect of learning styles on the writing process. *Computers & Education*, *73*, 60–71. https://doi.org/10.1016/j.compedu.2013.12.009.

Vehtari, A., Gelman, A., & Gabry, J. (2015). Pareto smoothed importance sampling. *arXiv preprint*, https://arxiv.org/abs/1507.02646

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4.

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage–Dickey method. *Cognitive Psychology*, *60*(3), 158–189. https://doi.org/10.1016/j.cogpsych.2009.12.001.

Wallot, S., & Grabowski, J. (2013). Typewriting dynamics: What distinguishes simple from complex writing tasks? *Ecological Psychology*, *25*(3), 267–280. https://doi.org/10.1080/10407413.2013.810512.

Wengelin, Å. (2006). Examining pauses in writing: Theory, methods and empirical data. In K. Sullivan & E. Lindgren (Eds.), *Computer key-stroke logging and writing: methods and applications (studies in writing)* (Vol. 18, pp. 107–130). Amsterdam: Elsevier.

Woong Yun, G., & Park, S.-Y. (2011). Selective posting: Willingness to post a message online. *Journal of Computer-Mediated Communication*, *16*(2), 201–227. https://doi.org/10.1111/j.1083-6101.2010.01533.x.

Xu, C. & Ding, Y. (2014). An exploratory study of pauses in computer-assisted EFL writing. *Language Learning & Technology*, 18 (3) 80–96. https://eric.ed.gov/?id=EJ1046527.

Zhang, M., Hao, J., Li, C., & Deane, P. (2016). Classification of Writing Patterns Using Keystroke Logs. In L. A. van der Ark, D. M. Bolt, W.-C. Wang, J. A. Douglas, & M. Wiberg (Eds.), *Quantitative psychology research: The 80th annual meeting of the psychometric society, Beijing, 2015* (pp. 299–314). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-38759-8_23.