Loughborough University

This item was submitted to Loughborough's Research Repository by the author.
Items in Figshare are protected by copyright, with all rights reserved, unless otherwise indicated.

Writing analytics across essay tasks with different cognitive load demands

PLEASE CITE THE PUBLISHED VERSION

https://2020conference.ascilite.org/proceedings/

PUBLISHER

Australasian Society for Computers in Learning in Tertiary Education (ASCILITE)

VERSION

VoR (Version of Record)

PUBLISHER STATEMENT

This is an Open Access Article. It is published by ASCILITE under the Creative Commons Attribution 4.0 International Licence (CC BY 4.0). Full details of this licence are available at: https://creativecommons.org/licenses/by/4.0/

LICENCE

CC BY 4.0

REPOSITORY RECORD

Oliveira, Eduardo, Rianne Conijn, Paula de Barba, Kelly Trezise, Menno van Zaanen, and Gregor Kennedy. 2020. “Writing Analytics Across Essay Tasks with Different Cognitive Load Demands”. Loughborough University. https://hdl.handle.net/2134/14331095.v1.

Writing analytics across essay tasks with different cognitive load demands

Eduardo Araujo Oliveira, Rianne Conijn, Paula De Barba, Kelly Trezise, Menno van Zaanen, Gregor Kennedy

The University of Melbourne, Eindhoven University of Technology, The University of Melbourne, Loughborough University, South African Centre for Digital Language, The University of Melbourne

Essay tasks are a widely used form of assessment in higher education. Writing analytics can assist with challenges related to using essay tasks at scale and to identifying different issues in academic integrity. In this paper, we combined two techniques to investigate how students' writing analytics varied across essay tasks with different cognitive load, considering both their typing behavior (i.e., writing process) and writing style (i.e., writing product). We also examined their relationship across these essay tasks. Findings showed that writing processes change across tasks with different cognitive load: when cognitive load increases, the interword intervals (indicator of planning and/or reviewing processes) increased, the burst length (indicator of translation processes) decreased, and the number of revisions per minute (indicator of reviewing processes) decreased. In contrast to the relation between the writing process and cognitive load, the relation between the writing product and cognitive load was found less clear. The results showed small and mixed effects of the tasks differing in cognitive load on the different writing product metrics. Hence, although the writing product follows from the writing process, the relation between cognitive load and the writing product and process appears to be less straightforward.

Keywords: writing analytics, learning analytics, stylometry, keystroke analysis, cognitive load.

Introduction

Writing analytics has the potential to improve teaching and learning in higher education. It can assist with challenges related to using essay writing at scale without over-burdening teaching staff (e.g., providing automated feedback; Abel, Kitto, Knight & Buckingham Shum, 2018) and identifying different issues in academic integrity writing (e.g., plagiarism, authorship verification, authorship attribution;(Abbasi and Chen, 2008; Calix et al., 2008; Schneider et al., 2017), for example. In order to fulfil this potential, we need to better understand how students' writing analytics vary across essay tasks with different complexity. Essay writing tasks can require students to achieve one or more learning objectives, such as verifying understanding and/or evaluating knowledge (Anderson, Krathwohl & Bloom, 2001). Different learning objectives have been found to demand different cognitive load from students; that is, the notion that a student's ability to perform a task depends on the cognitive demands of the task, and the student's working memory capacity available for task processing (Sweller, 1988). Previous research has found that essay tasks with different cognitive load impact different aspects of students' writing analytics, such as their typing behavior (e.g., Kormos, 2011; Van Waes, van Weijen & Leijten, 2014) and writing style (e.g., Deane, 2014). However, these aspects have been mainly investigated in separate studies, using approaches such as stylometry and keystroke dynamics, respectively (for an exception, see Brizan et al., 2015). In this paper, we combined these two techniques to investigate how students' writing analytics varied across essay tasks with different cognitive load, considering both their typing behavior (i.e., writing process) and writing style (i.e., writing product). We also examined their relationship across these essay tasks.

Background literature

Essay writing is a widely used form of assessment in higher education and it can be used to assess different learning objectives (Brizan et al., 2015). The Bloom taxonomy (Anderson et al., 2001) proposes six educational objectives: (1) remember, e.g., retrieval, (2) understand, e.g., interpret and explain, (3) apply, e.g., execute and implement, (4) analyse, e.g., organise and attribute, (5) evaluate, e.g., critique and make judgements, (6) create, e.g., generate and plan. These categories are thought to increasingly demand higher cognitive load from students; (Brizan et al., 2015). If the cognitive demands required for a given task exceed students' available working memory capacity, students' ability to perform the task will be affected. Students may take longer to

process information, use strategies that require less cognitive load, or make more errors. Previous research has found that such differences in cognitive load demands can be detected in essay writing using writing analytics.

Writing analytics has been defined as "the measurement and analysis of written texts for the purpose of understanding writing processes and products, in their educational contexts" (Buckingham Shum et al., 2016, p. 481). In the current study, we discriminate between data generated during the writing process and the writing product. Description and evaluation of writing processes have traditionally been somewhat subjective and difficult to quantify; relying on the think-aloud method and retrospective self-report interviews (Baaijen, Galbraith and Glooper, 2012; Leijten and Van Waes, 2013). Recently, keystroke data have been collected as students write, providing objective quantitative data and measuring writing processes on a finer-grained time-scale compared to think-aloud procedures (Baaijen, Galbraith and Glooper, 2012; Conijn, Van der Loo and Van Zaanen, 2018). For example, patterns of writing bursts and pauses have been associated with the three phases of writing: planning, translating, and reviewing (Baaijen, Galbraith and Glooper, 2012; Chenoweth and Hayes, 2001).

Stylometry is used to analyse static completed text (i.e., product), rather than the writing process. Stylometry is based on the linguistic style of the text produced by the author (Calix et al., 2008). The style of a completed text can be characterized by measuring a vast array of stylistic features, that includes lexical (e.g., word, sentence or character-based statistic variation such as vocabulary richness and word-length distributions), syntactic (e.g., function words, punctuation and part-of-speech), structural (e.g., text organization and layout, fonts, sizes and colours), content-specific (e.g., word n-grams), and idiosyncratic style markers (e.g., misspellings, grammatical mistakes and other usage anomalies) (Abbasi and Chen, 2008; Holmes and Kardos, 2003). Stylometry is often used for authorship identification, that is, modelling writing style to determine whether and how an author's identity can be inferred from their writing (Potthast et al., 2016). The use of stylometry for authorship identification assumes that an author's writing style is consistent and recognizable (Laramée, 2018)). Stylistic features are the attributes or writing-style markers that are the most effective discriminators of authorship. Over 1000 different style markers have been used in previous research on stylistic analysis, with no consensus on the best set (Rudman, 1997).

Keystroke dynamics and stylometry are a subset of a rich spectrum of cognition-centric behaviours that a typist exhibits during planning, translating, and reviewing text (Locklear et al., 2014). These approaches have been related to task load in different ways. First, a few studies have used keystroke and stylometry to predict writing context (e.g., stress). For example, Vizer, Zhou and Sears (2009) used keystrokes and stylometry to predict cognitive and physical stress. Keystroke features (including time per keystroke and mean pause duration) and stylometry features (including mean word length and lexical diversity) were identified as discriminators of stress conditions. Classification accuracy, indicating the ability of the models to predict stress from non-stress conditions, was up to 75%. In addition, another study indicated that stress, which is thought to disrupt cognitive functioning (Eysenck et al., 2007), is likely to change individuals' keystroke and stylometry patterns.

Second, several studies have examined the relationship between the type of task and keystroke or stylometry patterns. Two studies compared keystrokes across three tasks: copying from memory, copying from text, and generating text (e.g., describe the route from your home to the university) (Grabowski, 2008). These tasks were assumed to differ in cognitive demands. Copying from text was considered more demanding than copying from memory, because the former includes eye-hand coordination. In addition, text generation was considered to be more demanding than copying from text or memory, because the former includes some planning and formulation. These differing cognitive demands were reflected in the keystrokes. Text generation resulted in more pauses and fewer characters in the final text per total amount of keystrokes, compared to copying from memory or copying from text (Grabowski, 2008). In addition, the inter-keystroke intervals between words and within words was lower for copying from text, compared to copying from memory, but also compared to generating text (Grabowski, 2008).

Another study examined the relationship between type of task and both keystroke and stylometry metrics, through modulating the difficulty of text generation (Brizan et al., 2015). Task difficulty was based on Bloom's taxonomy, resulting in six tasks of varying cognitive load. In doing so, the study provided a measure of how writing processes and products vary with cognitive load. For example, the low cognitive load task asked participants to remember the reason why they joined a certain university, while the high cognitive load task asked them to plan their ideal class as if they were the teacher, including details about the subject and structure of tests (Brizan et al., 2015). Using a data-driven approach, 2098 keystroke features, 89 stylometry features, and

194 language production features were extracted. These features were used to classify the task (cognitive demand) with significantly more accuracy than the baseline.

Together, these studies suggest that varying task load can affect authors' writing process and writing outcomes, as reflected in keystroke patterns and stylometry. The significant implication here is that authorship identification/verification, or automatic assessment through keystroke or stylometry analysis, may be confounded if cognitive load alters keystroke or stylometry metrics to a significant degree. However, existing studies have largely focused on examining whether keystrokes and stylometry can predict low versus high cognitive load tasks. What is missing is an analysis of which aspects of the writing process and writing outcomes vary with cognitive load; i.e. what metrics change and in what way. Developing a characterization of the keystroke and stylometry changes associated with cognitive load offers two advantages: (1) understanding in what way the metrics obtained from the writing product and writing process are affected by essay tasks with different cognitive load demands, and (2) identification of the metrics that do not vary with cognitive load, so authorship identification/verification or automatic assessment analysis are not compromised with variations in essay task demands. This may in turn inform educators in the development of educational practices to improve writing for high cognitive load tasks (e.g., if the writing process of planning is affected, educators could increase the amount of class time spent on preparation before writing begins).

In the current study, we investigated the following research question: What is the impact of essay tasks with different cognitive load on writing processes (as measured by keystroke metrics) and on writing product (as measured by stylometry metrics)?

Method

Participants

The study was conducted at The University of Melbourne from 2017 to 2019. Participants were recruited via posters across the campus and provided informed consent (Ethics approval #1748727.1). The sample included a total of 46 students from four main disciplines: Engineering (24%, n=11), Commerce (24%, n=11), Arts (19.5%, n=9), Science (13%, n=6) and other (19.5%, n=9). Most participants were undergraduate students (70%, n=32), with 24 males (52%) and 22 females (48%). More than half of the participants were from a non-English speaking background (76%, n=35), and the majority of participants were right-handed (96%, n=44).

Procedure

In a computer laboratory, participants were asked to complete four activities using an Apple desktop computer and a QWERTY keyboard. The four activities were distributed over a period of 90 minutes (Figure 1). To account for possible effects of question ordering, two setups were used: one setup with increasing cognitive load, from low (1) to high (6) and one setup with decreasing cognitive load, from high (6) to low (1), as shown in Figure 1 The first 29 participants completed Setup 1, while the following 17 participants completed Setup 2. Prior to the first task, participants completed a pre-survey on demographic information. Participants completed four activities: Creative Work 1, Creative Work 2, Review, and Transcription. In the Creative Work 1 activity participants had 20 minutes to answer four open-ended questions requiring low to medium cognitive load (Q1, Q2, Q3, Q4; see Table 1). In the Creative Work 2 activity participants had 30 minutes to answer two open-ended questions requiring medium to high cognitive load (Q5, Q6; see Table 1). For the questions that required medium to high cognitive load, participants could consult two hardcopy supporting texts on the topic of university life. Participants then had a 10-minute break, where some snacks were provided. In the Review activity, participants had 10 minutes to review, edit and improve their answers from the Creative Work 2 activity (Q5a, Q6a; see Table 1). In Transcription activity participants were asked to transcribe one of the texts that was used as a support material during 'Creative Work 2' for 10 minutes (Q7).

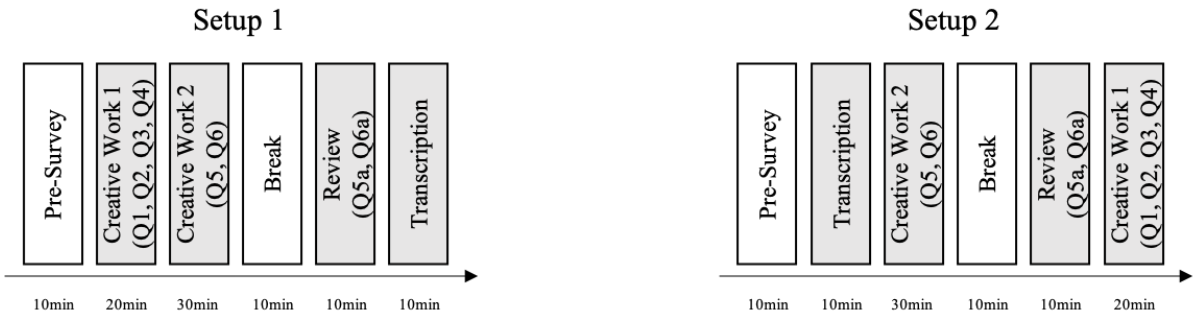


Figure 1: Breakdown of the tasks in the current study.

Transcription served as a reference task for the keystroke metrics. This study was part of a larger research design in which participants completed additional writing and survey tasks. These aspects of the research, however, are not the focus of the current paper.

Measures

Cognitive load was measured through six open-ended questions on the topic of university life (adapted from Brizan et al., 2015). Each question matched one of the six types of task in Bloom's Taxonomy: (1) remember, (2) understand, (3) apply, (4) analyse, (5) evaluate, and (6) create (Anderson et al., 2001). Table 1 lists the questions.

Table 1: List of questions and respective level of cognitive load.

ID	CL	Question
Q1	1	What made you decide to join this university?
Q2	2	What would you say has been the best class you have taken at this university and what did you enjoy about that class?
Q3	3	You are asked to complete a group assignment. It is important all students in the group contribute equally to the project. Come up with a plan for completing the group assignment, from research to class presentation.
Q4	4	Describe the similarities and differences between preparing a written assignment and preparing for a final exam.
Q5	5	A fellow university student spends a significant amount of their time worrying about their ability to complete their academic work, and becomes very concerned when they do not meet their grade expectations. In addition, they are concerned about financial pressures such as rent and textbook costs. Considering the texts you have received and the situation presented above, please answer the following question: Do you think the university should support this student improve their wellbeing? Why or why not?
Q6	6	[Using the scenario from Q5] Describe what advice you would provide to the student to help improve their wellbeing. What steps could they take?

Note. CL = Cognitive Load: expected demand based on Bloom's Taxonomy (Anderson et al., 2001), ranging from 1 = 'Low cognitive load demand' to 6 = 'High cognitive load demand'.

Participants' keystrokes and final texts were collected while answering the cognitive load questions and during the transcription task, using a writing software system. This system was developed as part of this research as a standalone Java application. The keystroke data recorded included: participant ID, question ID, key typed, key typed value, timestamp of each key press and release, hold time (milliseconds) and answer (text).

From the keystroke data, seven metrics were selected to use for the analysis of the following writing processes:

1. Planning and reviewing processes:
    a. Mean interword interval: the mean time (in milliseconds) between words, from key press of the last letter of the word until key press of the first letter of the subsequent word, without a pause. Here and in the following features, a pause is defined as an IKI longer than two SD from the mean IKI in the

transcription task (Deane, 2014).
   b. Pauses within words: Number of pauses within words, divided by the number of keystrokes. Here, a pause is calculated similarly as in the mean burst length.
   c. Pauses between words: Number of pauses between words, divided by the number of keystrokes.
2. Translation processes:
   a. Number of words per minute: the number of words typed, divided by the total time.
   b. Mean burst length: the average number of words per burst, where a burst is defined as a keystroke sequence that does not contain pauses or revision.
3. Reviewing processes:
   a. Revisions per minute: Number of revisions, divided by the total time. Here, a revision is a sequence of delete keys (e.g., backspace) that does not contain a pause.
   b. Keystrokes per character: Number of keystrokes, divided by the number of characters in the final product.

From the stylometry data, we used seven metrics across four dimensions to analyse the writing outcome:
1. Macro-organization dimension (calculated using TAACO 2.0; Crossley et al., 2019):
   a. Cohesion: Percentage of sentence linking connectives: the number of sentences linking connectives (e.g., therefore, consequently, moreover) divided by the total number of words.
   b. Coherence or semantic similarity: the semantic similarity in adjacent sentences (one sentence to the next sentence), calculated using word2vec vector space representations.
2. Complexity dimension (calculated using the syntactic complexity analyser; Lu & Ai, 2015):
   a. Mean length of T-unit: the total number of words divided by the number of T-units. T-units are minimal terminable units, which are the minimal units that could grammatically start with a capital and end with a period (Hunt, 1965). They are "grammatically capable of being considered a sentence" (Hunt, 1965, p. 21).
   b. Clause density: the number of main and subordinate clauses divided by the number of T-units.
3. Semantic complexity and readability dimension (calculated using the Natural Language Toolkit (NLTK) python library):
   a. Mean word frequency: the average word frequency, identified using the (case sensitive) word frequency per million words within the SubtlexUS Lexical Database of American English (Brysbaert & New, 2009). To account for misspelled words, the frequency of the most likely suggested word within the database was used. If no suggestions were made the word frequency was set to zero.
   b. Percentage of long words: the number of words with eight or more characters, divided by the total number of words.
4. Spelling dimension (calculated using the python library 'pyhunspell'; Latinier, 2019):
   a. Percentage of misspelled words: the number of misspelled words, divided by the number of words.

On average, the participants typed 114 words (SD = 87 words) per question, of which 100 words (SD = 79 words) ended up in the final product. The participants spent on average 11 minutes (SD = 12 minutes) per question. The descriptive statistics of the keystroke and stylometry metrics used for the analysis can be found in Table 2.

Analysis

Bayesian linear mixed effects models (BLMMs; Gelman et al., 2013; McElreath, 2018) were used to determine the effect of the questions with different levels of cognitive load on the keystroke and stylometry metrics. A Bayesian approach was chosen because it allows us to derive posterior probability distributions of the effect of the questions, which can be used to compare the effects of the questions across all keystroke and stylometry metrics. The BLMMs were implemented in R, using the R-package 'rstanarm'. The keystroke and stylometry metrics were used as dependent variables, and question (with the different levels of cognitive load) was added as fixed effect. In addition, as the effect of question on the metrics might differ across students, a by-participant slope for task was added. For the continuous dependent variables, we used linear models with log-normal distributions. For the ratio dependent variables, we used quasi-logit regressions, as these variables are bounded between 0 and 1 (see e.g., Donnelly & Verkuilen, 2017). Forward difference contrast coding was used to identify the differences in effect on the metrics, for every question compared to the reference level. With forward difference coding, one level of the categorical variable is compared with the next (adjacent) level of the categorical variable, i.e., question 1 is compared to question 2, question 2 is compared to question 3, and so on. The differences in effect were evaluated with 95% credible intervals, to determine the direction of the effect (for

more information see Kruschke, 2014; Nicenboim & Vasishth, 2016). In addition, we calculated the standardized effects (Cohen's d) from the posterior distributions of the models, to be able to compare the strength and the direction of the effects of all questions across keystroke and stylometry metrics.

Table 2: Descriptive statistics of the keystroke and stylometry metrics.

	Metric	Mean	SD
Keystroke	Mean interword interval (ms)	443	177
	Pauses within words	0.013	0.014
	Pauses between words	0.018	0.012
	Number of words per minute	21.0	14.1
	Mean burst length	4.69	4.01
	Revisions per minute	5.56	4.23
	Keystrokes per character	1.59	1.00
Stylometry	Percentage of sentence linking connectives	0.026	0.022
	Semantic similarity	0.367	0.199
	Mean length of T-unit	20.2	10.7
	Clause density	1.90	0.98
	Mean word frequency	5838	1810
	Percentage of long words	0.260	0.095
	Percentage of misspelled words	0.033	0.030

Results

Patterns in keystrokes

Figure 2a shows the standardized effects of the posterior distributions of all models on the keystroke metrics. Positive values indicate larger values for the question, compared to the subsequent question (question with higher cognitive demand); negative values indicate smaller values. Mean burst length, number of revisions per minute, and mean interword interval are most affected by the questions. This especially holds for question 4 compared to question 5, where the mean word interval is smaller, the mean burst length is larger, and the number of revisions is higher for question 4 compared to question 5. Thus, the participants had longer intervals between words, wrote shorter bursts, and made less revisions when the cognitive load increased from question 4 to question 5. The mean interword interval showed differences for only 2 of the questions. A negative effect was found for question 4 compared to question 5 and for question 1 compared to question 2. This indicates that the average time between words is smaller for questions 1 and 4 compared to the succeeding question with higher cognitive load (questions 2 and 5, respectively). The number of keystrokes per character in the final product only showed an effect for question 4 compared to question 5: less keystrokes per character were needed for question 4, compared to question

The effect of question was largest for the mean burst length and number of revisions per minute. Both keystroke metrics had a positive effect for question 4 compared to question 5. This means that the burst length is larger and the number of revisions per minute is higher for the question 4, compared to question with higher cognitive load (question 5). Thus, the writing is less fluent and less revisions are made with larger cognitive load. However, this effect is only visible for the fourth question compared to the fifth question, little difference is found in the mean burst length and number of revisions per minute for the other questions.

The results of the BLMMs credible intervals on the keystroke metrics include both positive and negative values for the effect of question on the pauses within words, the pauses between words, and the number of words per minute, indicating there is no stable effect (Figure 2a). This is also reflected in the small effect sizes (Cohen's d). Thus, the difference in cognitive demand measured through the six questions, has little effect on the pauses within words, the pauses between words, the number of words per minute, and the number of characters per keystroke.

Patterns in stylometry

Figure 2b shows the standardized effects of the posterior distributions of all models on the stylometry metrics. Cognitive load has little effect on the stylometry metrics. In addition, there is no clear set of consecutive questions for which the largest effects are found. As stated above, the percentage of long words and the percentage of misspelled words show effects in both directions, indicating that there is no clear linear relationship between these metrics and cognitive load. When comparing the outcomes of the keystroke metrics with the stylometry metrics (Figures 2 a and b) we can see that cognitive load has a larger effect on the writing process (keystroke metrics), compared to the writing product (stylometry metrics). For the clause density, percentage of long words, and percentage of misspelled words, small effects are found for multiple questions. The clause density is lower for question 1 compared to question 2, and for question 3 compared to question 4, indicating more clauses per T-unit when the cognitive load increases (but only for two questions). Interestingly, the percentage of long words and the percentage of misspelled words show differences in the directions of the effects across the questions.

The results of the BLMMs credible intervals on the stylometry metrics shows low effect sizes for almost all questions on the stylometry metrics, indicating that the cognitive load has limited effect on the stylometry metrics. For the effect of question on the mean length of a T-unit, all credible intervals include both positive and negative values, indicating that there is no stable effect of question on the number of words per T-unit. The percentage of sentence linking connectives, semantic similarity, and the mean word frequency show small effects for only one question, compared to the subsequent question. The percentage of linking connectives and the semantic similarity is lower for question 3 compared to question 4; and the mean word frequency is larger for question 1 compared to question 2, indicating more frequent words are used in question 1.

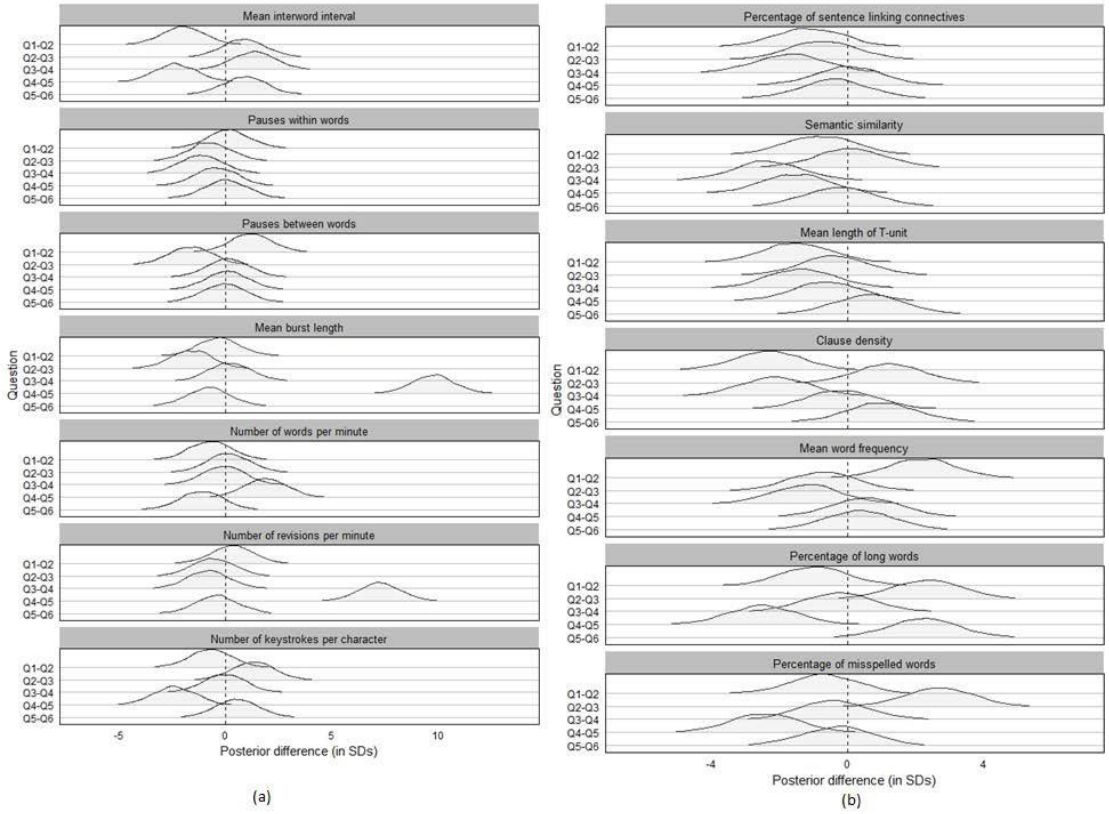


Figure 2: (a)Effect of question per keystroke metric, shown by the distributions of standardized posterior differences (Cohen's d). (b)Effect of question per stylometry metric, shown by the distributions of standardized posterior differences (Cohens' d).

Discussion

In this paper we presented how the writing processes, as shown by the keystroke metrics, differed with cognitive load. Cognitive load in planning and reviewing processes have been indicated by pause timings and interword

intervals (Medimorec & Risko, 2017), where longer and more frequent pauses are related to more effort (Grabowski, 2008; Van Waes et al., 2014). Our results only partly confirm this: the mean interword intervals increased for two subsequent questions, with increasing cognitive load (question 2-3 and question 4-5), however no effects of cognitive load were found for the pauses within and between words. Cognitive load in translation processes, also referred to as writing fluency, has been related to verbosity and burst length, where the number of words per burst and per minute is expected to decrease with cognitive load (Deane, 2014). This effect was only confirmed for the mean burst length for one subsequent question, where the burst length decreased when cognitive load increased from question 4 to question 5. Cognitive load in reviewing processes specifically have been indicated by the number of revisions per minute and the number of keystrokes per character in the final text. More experienced writers revise more and keep fewer of the proposed words in the final text (Chenoweth & Hayes, 2001; Choi, 2007). Hence, this would indicate that when cognitive load increases (less experience), the number of revisions decreases. This effect was found for question 5, compared to question 4. With higher cognitive load, less revisions were made per minute. However, the opposite was found for the number of keystrokes per character in the final product: with higher cognitive load more keystrokes are needed per character.

We also presented how the writing product, as shown by the stylometry metrics, differed with cognitive load. Here we considered three dimensions of written composition: macro-organization, complexity, and spelling (Wagner et al., 2011). Within macro-organization, higher cohesion and coherence have been related to higher writing proficiency (Kormos, 2011). Hence, we expected cohesion and coherence to decrease with cognitive load. However, the opposite was found: the percentage of sentence linking connectives and the semantic similarity increased with cognitive load, from question 3 to question 4. In addition, as higher working memory capacity is related to higher writing quality and complexity (Hoskyn & Swanson, 2003), we expected complexity to decrease with a higher cognitive load task. This effect was found for the percentage of long words long words (readability), where the number of long words decreased with cognitive load for questions 2-3 and 5-6. However, the number of long words increased with cognitive load for question 5 compared to question 4. In addition, no effects were found for the number of words per T-unit (syntactic complexity). Lastly, in contrast to what we expected, the number of clauses per T-unit increased with cognitive load (for question 1-2 and question 3-4) and the mean word frequency decreased with cognitive load (question 1-2). Thus, syntactic complexity and semantic complexity, respectively, increased with cognitive load. For spelling, we expected that the number of spelling errors would increase with cognitive load (Swanson & Berninger, 1996). However, mixed effects were found: the percentage of misspelled words indeed increased with cognitive load, from question 4 to 5, but decreased with cognitive load from question 2 to 3.

In summary, the results show an effect of cognitive load on the writing process and writing product for only certain pairs of subsequent questions and for some metrics. This might be because the differences in cognitive load between some consecutive questions are too small to be reflected in the keystroke and stylometry metrics. Previous work also found that more differences could be found between the extremes in terms of cognitive load (questions 1 and 6) rather than between subsequent questions (e.g., questions 1 and 2; Brizan et al., 2015). Even though the effects found are limited, they do provide some theoretical implications.

For the effect of cognitive load on the writing process, the largest effects were found for question 4 (analyse) compared to question 5 (evaluate). This suggests that these subsequent questions show the largest difference in cognitive load. This has important implications for understanding the differences in cognitive demand of the different levels of Bloom's taxonomy (Anderson et al., 2001), that is, it shows how distinct these levels are in terms of cognitive load. For the effect of cognitive load on the writing product there is no specific set of questions that shows the largest effects. In addition, several effects contradict our findings. For example, coherence and cohesion increased with cognitive load for questions 3 (plan) and 4 (analyse). This might be explained by the way students approached question 3. Without specifically prompting for a list, many bullet-point style responses were provided in question 3, which led to lower cohesion and coherence, whereas in question 4 students used a more descriptive style.

The study shows evidence that the writing process is affected by task cognitive load. While there were some effects of task cognitive load on writing product, these effects were moderate. However, given that the writing product is the outcome of the writing process, we would have expected similar effects of cognitive load on the writing process and product. This indicates that the relation between the writing product and writing process might not be as straightforward. Intuitively, certain aspects of the writing process are not visible in the writing product, e.g., if a sentence is rephrased 10 times, only shows the one final sentence in the writing product. Thus,

for the analysis of cognitive load in writing, and perhaps for the analysis of writing in general, we should not only focus on the writing product, but also analyse the writing process (in relation to the writing product).

The findings of this research have implications for educational practice. Educators and instructional designers could use keystroke and stylometry metrics to identify and compare the cognitive demands imposed by their chosen learning design. This is particularly relevant to educators setting open book exams in online environments. In addition, the relation between cognitive load and students' writing process and product may be used as a starting point for (personalized) feedback on students' writing processes, to improve their writing strategies. Moreover, this research has implications for authorship identification within academic institutions. The use of stylometry for authorship identification assumes that an author's writing style is consistent and recognizable (Laramée, 2018), much like a fingerprint. However, our findings show that some keystroke and stylometry measures commonly used in authorship identification studies can vary in response to the cognitive demands of the writing task. This raises questions as to whether the accuracy of keystroke and/or stylometry analysis for authorship identification is affected by the cognitive load of the writing task. If the ability to verify authorship is impaired, it suggests an important issue for academic integrity in educational institutions. Hence, it is important to analyse the accuracy of stylometry and keystroke metrics for authorship identification in educational contexts, where cognitive load can differ across tasks.

One limitation of the research is that cognitive load for each question was not measured, but rather, assumed based on previous research. Consequently, our claims on the relationship between cognitive load and writing patterns assumed that the different cognitive demands of each question would result in different cognitive loads (Anderson et al., 2001, Brizan et al., 2015). To test this, future work could measure the actual cognitive load, for example through interference in reaction times, or participants' self-reported cognitive effort.

Conclusion

This study shows that writing processes change across tasks with different cognitive load. The results showed that when cognitive load increases, the interword intervals (indicator of planning and/or reviewing processes) increased, the burst length (indicator of translation processes) decreased, and the number of revisions per minute (indicator of reviewing processes) decreased. These findings have important implications for the design and evaluation of writing tasks with different cognitive loads, as well as for authorship identification across tasks differing in cognitive load. In contrast to the relation between the writing process and cognitive load, the relation between the writing product and cognitive load was found less clear. The results showed small and mixed effects of the tasks differing in cognitive load on the different writing product metrics. Hence, although the writing product follows from the writing process, the relation between cognitive load and the writing product and process appears to be less straightforward.

References

Abbasi, A., & Chen, H. (2008). Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. *Transactions on Information Systems, 26*(2), 7.

Abel, S., Kitto, K., Knight, S., & Shum, S. B. (2018). Designing personalised, automated feedback to develop students' research writing skills. *Proceedings of the ASCILITE 2018 Open Oceans: Learning Without Borders*.

Anderson, L.W., Krathwohl, D.R., & Bloom, B.S. (2001). *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives.* Allyn & Bacon, Boston, MA.

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.

Brizan, D. G., Goodkind, A., Koch, P., Balagani, K., Phoha, V. V., & Rosenberg, A. (2015). Utilizing linguistically enhanced keystroke dynamics to predict typist cognition and demographics. *International Journal of Human-Computer Studies, 82*, 57–68. https://doi.org/10.1016/j.ijhcs.2015.04.005

Buckingham Shum, S., Knight, S., McNamara, D., Allen, L., Bektik, D., & Crossley, S. (2016). Critical perspectives on writing analytics. *Proceedings of the 6th International Conference on Learning Analytics & Knowledge*, pp. 481–483.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41*(4), 977–990.

Calix, K., Connors, M., Levy, D., Manzar, H., MCabe, G., &Westcott, S. (2008). Stylometry for e-mail

author identification and authentication. *Proceedings of the CSIS Research Day*, 1048–1054.
Chenoweth, N. A., & Hayes, J. R. (2001). Fluency in writing: Generating text in L1 and L2. *Written Communication, 18*(1), 80–98.
Choi, Y. H. (2007). On-line revision behaviors in EFL writing process. *English Teaching, 62*(4), 69–93.
Conijn, R., Van der Loo, J., & Van Zaanen, M. (2018). What's (not) in a Keystroke? Automatic Discovery of
Students' Writing Processes Using Keystroke Logging. *Proceedings of the 8th International Conference on Learning Analytics & Knowledge*. Sydney, Australia.
Crossley, S. A., Kyle, K., & Dascalu, M. (2019). The Tool for the Automatic Analysis of Cohesion 2.0: Integrating semantic similarity and text overlap. *Behavior Research Methods, 51*(1), 14–27. https://doi.org/10.3758/s13428-018-1142-4
Deane, P. (2014). Using writing process and product features to assess writing quality and explore how those features relate to other literacy tasks. *ETS Research Report Series, 2014*(1), 1–23. https://doi.org/10.1002/ets2.12002
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
Gelman, A., Stern, H.S., Carlin, J.B., Dunson, D.B., Vehtari, A., & Rubin, D.B. (2013). *Bayesian data analysis.* Chapman and Hall/CRC.
Grabowski, J. (2008). The internal structure of university students' keyboard skills. *Journal of Writing Research, 1*(1). https://doi.org/10.17239/jowr-2008.01.01.2
Holmes, D.I. & Kardos, J. (2003). Who was the author? An introduction to stylometry. *Chance, 16*(2), 5–8.
Hoskyn, M., & Swanson, H. L. (2003). The relationship between working memory and writing in younger and older adults. *Reading and Writing, 16*(8), 759–784. https://doi.org/10.1023/A:1027320226283
Hunt, K. W. (1965). Grammatical Structures Written at Three Grade Levels. *NCTE Research Report No. 3*. Retrieved from https://eric.ed.gov/?id=ED113735
Kormos, J. (2011). Task complexity and linguistic and discourse features of narrative writing performance. *Journal of Second Language Writing, 20*(2), 148–161.
Laramée, F. D. (2018). Introduction to stylometry with Python. *The Programming Historian, 7*. Retrieved from https://programminghistorian.org/en/lessons/introduction-to-stylometry-with-python
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.
Locklear, H., Govindarajan, S., Sitová, Z., Goodkind, A., Brizan, D.G., Rosenberg, A., Phoha, V.V., Gasti, P., & Balagani, K.S. (2014). Continuous authentication with cognition-centric text production and revision features. *Proceedings of the IEEE International Joint Conference*, 1–8.
Lockyer, L., Heathcote, E., & Dawson, S. (2013). Informing Pedagogical Action: Aligning Learning Analytics with Learning Design. *American Behavioral Scientist, 57(*10), 1439–1459. https://doi.org/10.1177/0002764213479367
Lu, X., & Ai, H. (2015). Syntactic complexity in college-level English writing: Differences among writers with diverse L1 backgrounds. *Journal of Second Language Writing, 29*, 16–27. https://doi.org/10.1016/j.jslw.2015.06.003
McElreath, R. (2018). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. 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
Nicenboim, B., & Vasishth, S. (2016). Statistical methods for linguistic research: Foundational Ideas—Part II.
Language and Linguistics Compass, 10(11), 591–613.
Paas, F. G., Van Merriënboer, J. J., & Adam, J. J. (1994). Measurement of cognitive load in instructional research. *Perceptual and Motor Skills, 79*(1), 419–430.
Potthast, M., Braun, S., Buz, T., Duffhauss, F., Friedrich, F., Gülzow, J.M., Köhler, J., Lötzsch, W., Müller, F., & Müller, M.E. (2016). Who wrote the web? Revisiting influential author identification research applicable to information retrieval. *Proceedings of the European Conference on Information Retrieval*, 393–407.
Rudman, J. (1997). The state of authorship attribution studies: Some problems and solutions. *Computers and the Humanities. 31*(4), 351–365.
Schneider, J., Bernstein, A., vom Brocke, J., Damevski, K., & Shepherd, D. C. (2017). Detecting plagiarism based on the creation process. *IEEE Transactions on Learning Technologies, 11*(3), 348–361. https://dx.doi.org/10.1109/TLT.2017.2720171

Swanson, H. L., & Berninger, V. W. (1996). Individual differences in children's working memory and writing skill. *Journal of Experimental Child Psychology, 63*(2), 358–385. https://doi.org/10.1006/jecp.1996.0054

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive science, 12*(2), 257– 285.

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

Vizer, L.M., Zhou, L., & Sears, A. (2009). Automated stress detection using keystroke and linguistic features: An exploratory study. *International Journal of Human-Computer Studies, 67*(10), 870–886.

Wagner, R.K., Puranik, C.S., Foorman, B., Foster, E., Wilson, L.G., Tschinkel, E., & Kantor, P.T. (2011).

Modeling the development of written language. *Reading and Writing, 24*(2), 203–220. https://doi.org/10.1007/s11145-010-9266-7

Oliveira, E., Conjin, R., de Barba, P., Trezise, K., Van Zaanen, M. & Kennedy, G. (2020). Writing Analytics Across Essay Tasks with Different Cognitive Load Demands. In S. Gregory, S. Warburton, & M. Parkes (Eds.), *ASCILITE's First Virtual Conference*. Proceedings ASCILITE 2020 in Armidale (pp. 60–70). https://doi.org/10.14742/ascilite2020.0121

Note: All published papers are refereed, having undergone a double-blind peer-review process.
The author(s) assign a Creative Commons by attribution licence enabling others to distribute, remix, tweak, and build upon their work, even commercially, as long as credit is given to the author(s) for the original creation.

© Oliveira, E., Conjin, R., de Barba, P., Trezise, K., Van Zaanen, M. & Kennedy, G. 2020