The KLiCKe Corpus: Keystroke Logging in Compositions for Knowledge Evaluation

Yu Tian[1], Scott Crossley[2] & Luuk Van Waes[3]

[1] Arizona State University | USA

[2] Vanderbilt University | USA

[3] University of Antwerp | Belgium

Abstract: Despite the growing interest in the dynamics of the writing process in writing research, publicly available large-scale corpora of keystroke logs have been rare. We introduce KLiCKe, a freely available collection of keystroke logs for around 5,000 argumentative texts written by adults in the United States. The KLiCKe corpus also includes human-rated holistic scores for the essays as well as writers' demographic details, their typing skills, and vocabulary knowledge. We describe our methods for constructing the corpus and present descriptives for different components of the corpus. To illustrate the use of the KLiCKe corpus, we report a study using a subset of the corpus to investigate whether keystroke features are associated with holistic writing quality for L1 and L2 writers. The study shows that higher writing scores are related to shorter pauses in general, shorter between-word pauses, lower proportion of deletions, higher proportion of insertions, and less process variance. The KLiCKe corpus provides a robust resource for researchers to study the dynamics of text production and revision that will help spur the development of process-oriented tools and methodologies in writing assessment and instruction.

Keywords: corpus, keystroke logging, writing quality

journal of WRITING RESEARCH

Tian, Y., Crossley, S. & Van Waes, L. (2025 - accepted for publication). The KLiCKe Corpus: Keystroke logging in compositions for knowledge evaluation. *Journal of Writing Research*, **, DOI: xx

Contact: Yu Tian, Arizona State University, 1151 S Forest Ave, Tempe, AZ. 85281. | Country – USA - ytian126@asu.edu - https://orcid.org/0000-0003-3009-3976

1. Introduction

Two important areas of writing research are writer behaviors during the writing process (e.g., planning, pausing, revising) and the characteristics of the written product (e.g., vocabulary, grammar, conventions, Conijn et al., 2022; Crossley et al., 2022; Lindgren & Sullivan, 2019; Van Waes & Leijten, 2015; Wengelin, 2006). Information about the writing process can provide valuable insights into writers' behaviors and the cognitive activities that lead to differences in the text writers generate. Understanding the writing process is important for investigating various issues including writing engagement, writing difficulties, and planning and revision strategies, all of which can be used to provide process-based evaluation and feedback to writers and teachers (Bowen et al., 2022; Sinharay et al., 2019; Vandermeulen et al., 2020, 2023).

Since its inception, keystroke logging has become a major observational tool in writing process research, particularly in the contexts of timed writing tasks (Almond et al., 2012; Barkaoui, 2016; Sinharay et al., 2019), academic writing (Choi & Deane, 2020), professional writing (Leijten et al., 2013), and language translation (Dragsted & Carl, 2013). Recent studies have revealed links between keystroke features obtained from writing processes and characteristics of finalized products such as lexical diversity (Medimorec & Risko, 2016), text cohesion (Tian et al., 2021), argumentation (Tian et al., 2024), and writing quality (Allen et al., 2016; Choi & Deane, 2020; Deane & Zhang, 2015), which provides insights into how writer behaviors and strategies may affect the written product.

Keystroke logging data also show promise in the development of automated writing evaluation (AWE) tools and other writing platforms. For example, Inputlog (Leijten & Van Waes, 2013), a widely used keystroke logging program, has a built-in report function to facilitate feedback to students. The function automatically generates a report that features a selected list of process-based writing aspects including time characteristics, process description, pausing, revision, source use, typing characteristics, and process and fluency graphs. Research has shown that these process-oriented features positively impact writing interventions for high school students (Vandermeulen et al., 2023). Another tool that leverages keystroke logging data to provide writing process feedback is Cywrite (Chukharev-Hudilainen, 2019), a web-based low-feature text editor with built-in keystroke logging, eye tracking, and automated writing evaluation. Cywrite triangulates learners' linguistic data with keystroke logging information to infer their cognitive processes. The system then provides real-time feedback to support learners' text production when they are struggling in writing (e.g., they exhibit a pause longer than 10 seconds).

Given the value of keystroke logging data in writing research, assessment, and pedagogy, a large-scale, publicly available keystroke-logged corpus could significantly advance the field. Such a resource would catalyze research, enhance our understanding of the writing process, facilitate the development of more robust writing assessment

methodologies, and foster process-oriented pedagogical interventions in writing instruction. To our knowledge, however, publicly available corpora of this kind are rare.

In this paper, we introduce the Keystroke Logging in Compositions for Knowledge Evaluation (KLiCKe) corpus, a freely available collection of around 5,000 argumentative essays with detailed keystroke logs of writing process information generated by adult writers in the United States. Keystroke data in KLiCKe corpus was collected through crowdsourcing using a web-based keystroke logging program that unobtrusively recorded every keystroke and mouse action, along with time stamps and cursor position information. In addition, all the essays were human-scored for holistic writing quality based on a standardized writing evaluation rubric. The corpus also includes writers' demographic information (e.g., gender, age, native language, education) as well as results from typing skills and vocabulary knowledge tests. To illustrate the use of the KLiCKe corpus, we report a study investigating whether keystroke features are associated with holistic writing quality for L1 and L2 writers.

1.1 Keystroke Logging

To analyze the dynamic nature of the writing process, researchers have developed and used a variety of research methods (Mackey & Gass, 2015). Traditional methods such as think-aloud protocols (e.g., Hayes & Flower, 1981; van den Bergh & Rijlaarsdam, 2007) and retrospective reports (e.g., Lindgren & Sullivan, 2003; Schumacher et al., 1984) have proven effective in uncovering information that links to writers' cognitive processes in text production but come with reliability and validity concerns (Janssen et al., 2013). Think-aloud protocols, for instance, require writers to verbalize their thoughts as they write, potentially disrupting their cognitive processes during writing (Russo et al., 1989). Retrospective reports, although less intrusive, rely on writers' memory, which may lead to inaccuracies in reconstructing experiences (Wengelin et al., 2019). Additionally, screen-capturing has also been used as a non-intrusive method for observing writing processes, but it is more suited for providing pedagogical feedback rather than in-depth quantitative analysis of writers' behavioral patterns (e.g., Hamel & Séror, 2016; Stannard, 2019).

As typing has become prevalent in text production, more attention has been drawn to keystroke logging as an observational tool in writing research. Keystroke logging has generally been implemented by activating a program that records keystrokes and mouse clicks or movements with time stamps during text production (Leijten & Van Waes, 2020). As an observational tool in writing research, keystroke logging offers advantages over the traditional methods despite the challenges in aligning certain keystroke measures with specific cognitive processes in writing (Galbraith & Baaijen, 2019). Firstly, it captures the temporal details of every keystroke and mouse movement in the unfolding writing processes, both unobtrusively and ecologically (Lindgren & Sullivan, 2019). The large amount of fine-grained data collected via keystroke logging allows for in-depth analyses of various writing behaviors such as pauses and revisions in composing. Secondly, keystroke logging is more scalable compared to other observational methods

such as screen recording and think-aloud protocols. To date, a number of laboratory-based keystroke logging programs have been developed, among which Scriptlog (Strumqvist et al., 2006), Translog (Carl, 2012), Inputlog (Leijten & Van Waes, 2013), and GenoGraphiX-LOG (Caporossi, Leblay, & Usoof, 2023) are most notable. These programs were developed with different focuses. Specifically, Scriptlog was designed for use in highly controlled experimental research, integrating also eyetracking (Wengelin et al., 2024). Translog was developed for translation studies. Inputlog was devised to capture and analyze writing activities (including source use) in both ecological and experimental settings (Leijten & Van Waes, 2020). GenoGraphiX-LOG is a web-based logger and was designed to log, analyze, and visualize writing process data in three writing contexts: free-writing, translating, and editing translation. In general, these keystroke logging programs record every keystroke operation such as insert, delete, cut, paste, and replace as well as every cursor and mouse movement as writers produce text on a computer. The logged events are time coded to indicate when the events occur and how long they last. The text position information for these events can also be captured, allowing for finer-grained analyses of writing behaviors, such as pauses at the boundaries of different linguistic units and revisions.

1.2 Keystroke Features and Writing Quality

Keystroke logging enables the quantification of various temporal features of the writing process, allowing researchers to operationalize behaviors such as pausing, revising, and language bursting (i.e., fluently producing a stretch of text with no long pauses or revisions; Kaufer et al., 1986). By using temporal features derived from keystroke logging, researchers can probe the relationship between these features and writing performance in educational contexts.

Recent research has evaluated the links between writing process features and writing performance in the K-12 assessment context and has delineated a set of keystroke features that are predictive of writing quality. For instance, total writing time, the number of keystrokes, and typing speed, which are measures of general writing fluency and efforts, are related to writing quality among young writers (Sinharay et al., 2019; Zhang et al., 2016). Regarding pauses, Deane (2014) found that higher-performing students in writing assessments demonstrated shorter pauses at character, word, and sentence boundaries. Likewise, a study conducted by Zhang et al. (2016) suggested that shorter pre-writing pauses under a certain timed-writing test condition indicated an adequate understanding of the task requirements, greater familiarity with the writing topic, and better task planning. Additionally, burst lengths and variations have been reported to show predictive power of essay quality in timed writing assessments (e.g., Deane & Zhang, 2015). Deane and Quinlan (2010) and Deane (2014) documented that stronger writers produced text more efficiently in longer burst spans. Sinharay et al. (2019) also found that both the number of bursts and the burst length predicted higher essay scores. In terms

of revising behaviors, studies have generally shown that students of higher writing performance tended to make more revisions in their text production (e.g., Deane, 2014).

Process features derived from keystroke logs have also been reported to be associated with writing performance for adult writers in recent years. Allen and her colleagues (2016) analyzed the keystroke logs collected from 126 undergraduate student writers in argumentative writing sessions using an intelligent tutoring system. The study reported that keystroke indices accounted for 76% of the variance in essay quality. In particular, the results indicated that stronger writers demonstrated a higher and more consistent production rate (more keystrokes) over the course of the writing session. Moreover, these writers' text production processes were also characterized by shorter pause times. Similar results have been reported in research conducted in English as second language (ESL) and foreign language (EFL) writing assessment contexts. Révész, Michel, and Lee (2017) studied the online writing behaviors of 30 Mandarin users of L2 English at a UK university in a standardized argumentative writing test and found that better performing L2 writers produced text with higher fluency and less frequent pausing within words. Choi and Deane (2020) evaluated the predictive potential of process features extracted from keystroke logs of adult EFL writers (N = 798) in an assessment context. They utilized the keystroke features to construct models to predict human-rated writing quality scores. The results showed that keystroke features significantly improved the predictive power of the models over the baseline, substantiating the associations between L2 keystroke features and writing quality in assessment contexts. Similar to previous L1 writing research, Choi and Deane also concluded that L2 writers who scored higher in argumentative writing tests tended to produce text with a higher fluency (more keystrokes recorded) and shorter pauses. Xu (2018) investigated the online revisions of 57 Chinese EFL writers and analyzed the relationship between their revising behaviors and writing quality. The analyses revealed that less-skilled L2 writers revised more frequently on smaller scopes during text production while more-skilled L2 writers revised more frequently on larger scopes after the completion of the main text.

One area of contention within the examination of keystroke characteristics in relation to writing proficiency involves the stability and generalizability of these process features across writing tasks or genres with different groups (e.g., age, gender, socioeconomic status) of writers. Studies have shown varied keystroke characteristics among writers when they responded to tasks of different complexity (Jung, 2017) or different writing genres (Olive et al., 2009). Keystroke features have also been associated with writers' gender (Guo et al., 2019; Zhu et al., 2019), ethnicity (Guo et al, 2019), working memory capacity (Ransdell et al., 2001), and language proficiency (Barkaoui, 2019; Van Waes & Leijten, 2015; Zhu et al., 2019). Several studies have investigated the stability of keystroke metrics in predicting writing quality across different assessment occasions. Deane and Zhang (2015), for example, examined the feasibility of modeling writing quality using the keystroke logs collected from a large group of adolescent students in the U.S. who wrote in six writing test forms that covered two genres (argumentative essay and written recommendation), each containing three different topics. The results showed moderate

to strong prediction of human-rated essay scores by a set of stable keystroke features generalizable across writing genres and prompts, although keystroke features vary considerably across different writing test occasions. Choi and Deane (2020) evaluated the stability of keystroke features of adult EFL learners in an assessment context from multiple perspectives: across different time points within a response and across responses to different tasks. The study found that most keystroke features were stable and exhibited meaningful correlations with writing quality, although a large variance was detected in terms of both within-response and within-person stability. Although limited in scope and scale, these studies provide empirical support for the use of keystroke features in measuring writers' fundamental text production skills and modeling writing performance.

1.3 Keystroke Logging Databases

One hurdle in better understanding the role that keystroke logs play in explaining the writing process and its relation to real-world applications is the lack of robust and large-scale open keystroke logging databases. One of the plausible reasons for this lack is that the dominant method for keystroke logging in writing research has been laboratory-based, which often requires the installation of specialized keystroke logging software (e.g., Inputlog) on computers prior to data collection (e.g., Barkaoui, 2019; Rossetti & Van Waes, 2022). This method is effective for research that involves experiments with highly controlled conditions, but may not be well suited for large-scale research where keystroke information needs to be obtained from a large pool of participants. Although researchers at Educational Testing Service (ETS) have conducted large-scale investigations of the writing process in assessment settings using keystroke logs collected from multiple test centers across the world, these datasets are generally not freely available.

There are a few existing keystroke-logged writing datasets that are publicly accessible. For instance, the LIFT project, conducted by Vandermeulen et al. (2020), involved 617 Dutch students from grades 10, 11, and 12 across 43 schools, who wrote various source-based texts. This dataset includes writing process data logged and analyzed via Inputlog, including production, pausing, revision, and source use. Additionally, it encompasses text quality scores provided by three raters for each text, and basic participant information such as gender, age, grade, and native language. The dataset also includes the type of writing task (whether argumentative or informative), the topic, and source complexity. Another example is Pro-Text (Miletić et al., 2022), an annotated corpus of keystroke logs with 202K tokens written in French. Its keystroke information was recorded using Inputlog (on Windows) and Scriptlog (on macOS). There are five sub-corpora in Pro-Text: Academic (26 mini-theses by MA students), Professional (10 reports by social workers), Experimental (165 essays by BA students in experimental conditions), Children (183 narrative texts and essays by schoolchildren), and Translation (38 original and translated texts by BA students). Some writing researchers have also released their keystroke logging

data as open-source materials (e.g., Mucoz Martín & Apfelthaler, 2022; Vandermeulen et al., 2020), although these datasets are generally small (e.g., < 100 participants).

1.4 Current Study

This paper introduces the Keystroke Logging in Compositions for Knowledge Evaluation (KLiCKe) corpus, a large-scale corpus of argumentative texts written in English with keystroke logs related to the text production process. The main goal of the KLiCKe corpus is to advance process-oriented writing research, pedagogy, and the development of automated diagnostic systems that provide timely feedback on learners' writing processes.

2. Data Collection

2.1 Procedure and Apparatus

The data for the KLiCKe corpus were collected from January through November 2022 via Amazon Mechanical Turk (MTurk), an online crowdsourcing platform. We hired workers (referred to as "Turkers") from MTurk who met three threshold qualifications: 1) be at least 18 years old; 2) be currently living in the United States; and 3) have completed at least 50 MTurk tasks with an overall approval rate of at least 98% by experimenters on the platform. We invited the Turkers to log onto a project-specific website to complete several tasks, including a demographic survey, typing tests, an argumentative writing task, and a vocabulary knowledge test. Prior to initiating these tasks, participants were presented with an informed consent form outlining the study's objectives and procedures. They were also instructed to complete the tasks in a quiet and distraction-free environment to ensure the quality and reliability of the collected data. During data collection, Turkers were required to use only computers with a keyboard. Their keystroke activities during the typing tests and the argumentative writing task were recorded using a built-in keystroke logging program that captured timing and cursor position information for every keystroke and mouse operation. Approximately 10,000 Turkers attempted to work on our project through MTurk, but not all completed the work. Each successful Turker data collection lasted around 40–50 minutes. All Turkers were paid a $0.25 reward, and successful Turkers were paid a $11.75 bonus upon completing all tasks as per the instructions on the website.

2.1.1 Demographic survey

We administered the demographic survey through Qualtrics Research Suite software (Qualtrics, 2022). The survey included questions about writers' demographic information, including gender, age, citizenship, race/ethnicity, education, and language background. For non-native English writers, the survey asked additional questions about their native languages and English learning experiences, including the age at which they began learning English, the number of years they have studied English, and the number

of years they have lived in the United States. See Appendix A for a full list of the questions used in the demographic survey.

2.1.2 Typing tests

Typing tests to measure workers' typing skills were adapted from the Inputlog copy task (Van Waes et al., 2019). The tests included a tapping test that measured the fastest motor speed of pressing two keys (e.g., "d" + "k") alternatively in 15 seconds, a sentence copying test that measured typing skills related to repetitively copying a short sentence comprised of high frequency words (e.g., "the cat was sleeping under the apple tree") in 30 seconds, a series of three-word combination copying tests that contained four word combinations and measured the speed of copying a set of bigrams of low and high frequencies (e.g., "five interesting questions", "five important behaviors", "some awkward zigzags"), and a consonant copying test that assessed typing skills in a non-word context (e.g., "tjxggl"). The typing tests took writers around 5 minutes to complete on average.

2.1.3 Argumentative writing task

In the argumentative writing task, writers were asked to write an argumentative essay within a 30-minute timeframe in response to a writing prompt adapted from a retired Scholastic Assessment Test (SAT) taken by high school students attempting to enter post-secondary institutions in the United States. To control for potential prompt effects, four SAT-based writing prompts were used, and each writer was randomly assigned one prompt. Appendix B presents the four prompts used. Prior to the writing task, writers were given instructions about important components of an argumentative essay (e.g., introduction, position, reasons and evidence, counterarguments and rebuttals, and conclusion) along with descriptions of their functions in argumentation. The instructions also introduced a set of suggestions for the writing task. These included that writers should write essays of at least 200 words using at least three paragraphs, and that they should not use any online or offline reference materials. To help writers stay focused on the task during writing and to track behavior, the writing task page issued warnings whenever the writer was detected to be inactive for more than 2 minutes or moved to a new window in the process of writing. A screenshot of the writing task page is presented in Figure 1.

2.1.4 Vocabulary knowledge test

To assess writers' vocabulary knowledge, we adapted the Lexical Test for Advanced Learners of English (LexTALE), a time-efficient test that has been validated as a reliable predictor of English vocabulary knowledge (Lemhufer & Broersma, 2012). In this test, writers were presented with 60 trials. In each trial, a string of letters was shown on the screen (see Figure 2). Writers were asked to decide whether the string was an existing English word or not by clicking "Yes" (word) or "No" (nonword). A total of 40 English words and 20 nonwords were used for the trials. During each trial, workers were given

as much time as needed for their decision. The vocabulary test took around 3–4 minutes to complete on average. The results were stored in a tabular format (see Table 1) where the ***Word*** column stores the words/nonwords used in the test, the ***Response*** column presents writers' answers, and the ***Key*** column shows the correct answers. In both ***Response*** and ***Key*** columns, the value of "1" denotes "Yes," while "0" signifies "No."

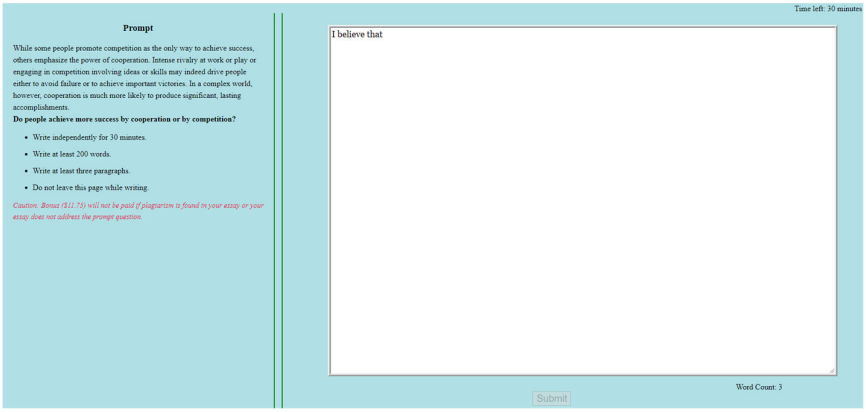

Figure 1. A screenshot of the writing task web page.

Note. The webpage interface features a randomly assigned prompt and task rules on the left, with a text area for essay composition on the right. A countdown timer in the upper-right corner tracks remaining time, and the word count appears in the lower-left. The "Submit" button is disabled until after 30 minutes.

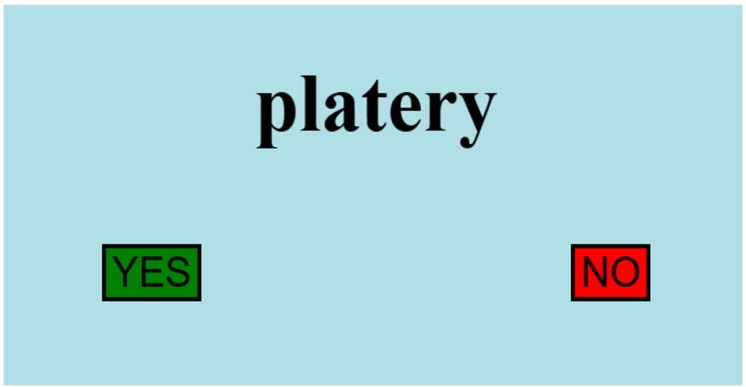

Figure 2. A screenshot of the adapted LexTALE.

Table 1. An example dataframe for vocabulary knowledge results

Word	Response	Key
platery	1	0
denial	1	1
generic	0	1
mensible	1	0
scornful	0	1
stoutly	1	1
ablaze	0	1
kermshaw	1	0
moonlit	1	1
lofty	1	1

2.1.5 Keystroke logging program

To collect Turkers' keystroke information during the typing tests and the argumentative writing task, a keystroke logging program was written in JavaScript and was hooked to the text input area on the webpage. The program unobtrusively recorded every keystroke and mouse activity along with relevant timing and cursor position information (for instances of text selection, the default cursor position was recorded as the endpoint of the selection). Additionally, the program concurrently analyzes writers' typing behaviors (e.g., undo/redo actions, text dragging, and mouse navigation), identifies operation types (e.g., input, delete, paste, replace), and reported text changes in the writing process. Table 2 provides an example output of keystroke logging information reported by the program.

Each entry includes the following fields:

- *Event ID.* Indexes keyboard and mouse operations in chronological order.
- *Down Time*. Records the time (in milliseconds) when a key or the mouse was pressed.
- *Up Time.* Marks the release time of the event.
- *Pause Time.* Represents the time between two consecutive key presses (i.e. *Press down^{y} - Press down^{x}).*
- *Action Time.* Measures the action time or duration of the operation (i.e., *Up Time - Down Time*) (Note that the operation time for Control, Alt, and Shift keys was set to 0 to facilitate the logging of non-modifier keys).
- *Position.* Registers cursor position information to help keep track of the location of the leading edge.
- *Word Count.* Displays the accumulated numbers of words typed.
- *Text Change.* Captures exact modifications made to the text.
- *Activity.* Classifies the nature of the changes (e.g., Input, Remove/Cut, Replace).

These records collectively provide a comprehensive view of the writer's keyboard and mouse interactions during writing. When analyzed together, they enable the reconstruction of text production processes, including complex actions such as copy-pasting, text dragging, and undoing changes.

Table 2. An example dataframe of keystroke logging information

Event ID	Down Time	Up Time	Action Time	Event	Position	Pause Time	Word Count	Text Change	Activity
1	746	791	45	Leftclick	0	0	0	NoChange	Nonproduction
2	19948	19948	0	Shift	0	19202	0	NoChange	Nonproduction
3	20210	20321	111	I	1	262	1	I	Input
4	20369	20441	72	Space	2	159	1	Space	Input
5	20417	20497	80	b	3	48	2	b	Input
6	20601	20730	129	e	4	184	2	e	Input
7	20658	20809	151	l	5	57	2	l	Input
8	20769	20921	152	i	6	111	2	i	Input
9	20849	21001	152	e	7	80	2	e	Input
10	20937	21097	160	v	8	88	2	v	Input
11	21034	21185	151	Space	9	97	2	Space	Input
12	21105	21257	152	Backspace	8	71	2	Space	Remove/Cut
13	21209	21321	112	e	9	104	2	e	Input
14	21369	21497	128	Space	10	160	2	Space	Input
15	21425	21506	81	t	11	56	3	t	Input
16	21441	21513	72	h	12	16	3	h	Input
17	21585	21745	160	a	13	144	3	a	Input
18	21633	21753	120	t	14	48	3	t	Input
19	21689	21825	136	Space	15	56	3	Space	Input

The web-based keystroke logging program was rigorously tested to ensure consistent and reliable performance across major operating systems (e.g., Windows, macOS, Linux) and commonly used web browsers (e.g., Chrome, Firefox, Edge). To assess its temporal accuracy for the KLiCKe corpus, we performed a series of tests based on the research protocol established by Frid et al. (2012). The results indicated that the web-based program achieved accuracy comparable to that of Inputlog 9 across different browsers, including Google Chrome and Mozilla Firefox. For a detailed description of the test procedures and results, we refer to Appendix C.

2.2 Data Extraction, Cleansing, and Transformation

All the data except the demographic information (collected using Qualtrics) were ingested into a MongoDB Atlas database (https://www.mongodb.com/atlas/database). We built a data pipeline to first extract the data from the database and then converted the data into target formats to facilitate further cleaning and analysis. Specifically, the keystroke information in the argumentative writing task and the typing tests as well as the results in the vocabulary knowledge test were reformatted into .csv files. We also

extracted from the database the content of the essays submitted by the writers during the argumentative writing task and output them into .txt files.

We implemented strict procedures in checking and cleansing the data. First, each essay was checked to ensure that it was devoid of plagiarism and met the writing requirements. To identify instances of plagiarism within the essays, we used Grammarly's plagiarism detection tool (Grammarly, 2022) which provided an assessment indicating the proportion of text deemed to be plagiarized, along with corresponding source references (note that the data was collected before the release of ChatGPT). Essays demonstrating a plagiarism ratio exceeding 10% underwent further scrutiny via manual verification to corroborate Grammarly's report. Furthermore, each essay was eyeballed to check whether it addressed the assigned prompt and was argumentative in nature. Submissions displaying overly generic responses or repetitive sentence patterns indicative of non-human output were flagged for collective review. To ensure data quality and integrity, essays found to contain plagiarism, essays that failed to meet the writing requirements, or essays that were likely generated by a bot were excluded from the dataset. Corresponding keystroke logs collected during the argumentative writing task were also removed. Note that 40 essays were below 200 words but were still retained in the corpus because the keystroke logs from these struggling writers may be of value. Second, we checked the keystroke data for any technical errors and anomalies. Specifically, we processed the keystroke logs through the pipeline to detect any missing information, unrecognized keystrokes, and irregular patterns (e.g., exceedingly long pauses). Additionally, instances where substantial text appeared to have been pasted from external sources were detected, as these could indicate potentially dishonest writing behavior. Any files that were deemed flawed based on these criteria were excluded from the dataset. Third, to safeguard the integrity and reliability of the dataset, we identified and eliminated both duplicated data entries (i.e., instances of replicated data generated by the same author) and incomplete data entries (characterized by missing information in demographic details and argumentative writing data). Among all collected essays, approximately 32% were found to be plagiarized. Out of the un-plagiarized texts, 14% did not meet the writing requirements. In addition, about 2% of keystroke logs contained one of the issues stated above. Lastly, around 1% of the Turkers did not provide full demographic information or did not complete the writing task. After checking and cleansing the data, we retained data entries from 4,992 MTurk workers.

In addition to the .csv files of the keystroke data for the argumentative writing task and the typing tests, we also converted the keystroke data into .idfx files using a custom JavaScript script. This script adhered to the technical standards outlined in Van Waes et al. (2012). Specifically, it reformatted each row of the .csv files by embedding the data within XML tags and attributes that conform to the Inputlog format specifications. The accuracy of this data transformation was validated through comparative analyses. First, we simultaneously recorded a series of writing processes using Inputlog and our web-based program. We then analyzed the generated .idfx files using Inputlog's general

analysis module. By comparing the logs produced by Inputlog with those generated by our reformatting process, we confirmed that the transformation maintained the integrity and precision of the keystroke data. We did this transformation in hope that users without much coding expertise can analyze the keystroke data by simply importing the .idfx files into Inputlog (version 9 or higher) and producing keystroke metrics using its built-in analysis modules such as summary, pause, revision, and fluency analyses.

2.3 Essay Quality Scoring

The essays were also scored by trained raters for overall writing quality using a holistic, six-point grading scale commonly used in assessing SAT essays (see Appendix D). The holistic rubric evaluated writing quality on multiple dimensions, including writers' development of a point of view on the issue, evidence of critical thinking, use of appropriate examples, accurate and adept use of language, the variety of sentence structures, errors in grammar and mechanics as well as text organization and coherence.

We hired and trained thirteen human raters to assess the essays for overall writing quality. The raters were graduate students majoring in either English or applied linguistics. All of them had at least two years of experience teaching English composition at the university level. All raters went through at least three rounds of training sessions before they scored the essays independently. The training included an introduction about essay collection methods, the holistic rubric, and the prompts, as well as a discussion of avoiding potential biases in essay scoring. In each session, raters independently scored a batch of essays before they met to discuss the differences in their scores. A total of 60 practice essays were used for training purposes. These practice essays covered the same topics but were sampled from a different dataset.

During training, we calculated weighted Cohen's Kappa to measure inter-rater reliability among the raters. Raters completed the training and began independently scoring the essays only after achieving an acceptable agreement level, with a Cohen's Kappa of at least 0.600 (Cohen, 1960). The allocation of essays among the raters post-training was randomized, with each essay being evaluated by at least two raters. The Initial Cohen's Kappa obtained for the entire dataset was $k = 0.601$, $p < .001$, indicating substantial overall agreement between the raters. If the score difference between two raters was two points or more, they adjudicated the scores through discussion. If agreement was not reached, the score was not changed. The Cohen's Kappa after adjudication was $k = 0.759$, $p < .001$, reflecting a marked improvement in inter-rater reliability. This improvement was likely due to adjudication discussions, which helped address potential oversights or biases in the initial ratings. The average of the adjudicated holistic scores from the raters was calculated for each essay and used in the final dataset to represent writing quality.

3. An Overview of the KLiCKe Corpus

The final KLiCKe corpus comprises 4,992 argumentative texts with the corresponding keystroke logs (in both .csv and .idfx formats) of the writing process and holistic scores for writing quality. Additionally, it includes the authors' demographic details, keystroke logging data from the typing tests (represented in .csv and .idfx files), and the vocabulary knowledge test results. Below, we present descriptive statistics for the various components of the KLiCKe corpus.

3.1 Demographic Details

The demographic information for the MTurk writers whose data entries were included in the final corpus is presented in Table 3. As shown, more female writers (58.39%) than male writers (40.1%) participated in the study. Writers' ages ranged from 18 to over 61 years, with a mean of 37.82 (SD = 12.08). The majority of the writers were white, had a college degree, and spoke English as their native language. Non-native English speakers represented diverse linguistic backgrounds, with Spanish being the most common, followed by French and Chinese. On average, non-native speakers had studied English for 16.34 years (SD = 9.57) at the time of data collection.

Table 3. Demographic information for MTurk workers

Items	**Participants (*N* = 4992)**
Gender	Female (2915); Male (2002); Non-binary/third gender (63); Prefer not to say (2)
Age	18–30 (1597); 31–40 (1714); 41–50 (846); 51–60 (527); 61+ (308)
Ethnicity	White (3809); Black or African American (389); Asian (299); Hispanic or Latino (282); American Indian or Alaska Native (77); Native Hawaiian or Pacific Islander (8); Multiple ethnicity/other (5)
Education (highest degree completed)	Less than high school diploma (38); High school diploma or the equivalent (415); Some college, but no degree (845); Trade/technical/vocational training (145); Associate degree (450); Bachelor's degree (2172); Master's Degree (768); Professional Degree (59); Doctorate (93); Unidentified (5)
Native language	English (4650); Spanish (121); French (30); Chinese (27); Tamil (22); Hindi (15); German (14); Arabic (9); Russian (8); Korean (6); Turkish (6); Urdu (6); Japanese (5); Portuguese (5); Swahili (5); Tagalog (5); Bengali (4); Polish (4); Vietnamese (4); Filipino (2); Greek (2); Haitian (2); Hebrew (2); Indonesian (2); Nepalese (2); Romanian (2); Serbian (2); Telugu (2); Zulu (2); Other (26)

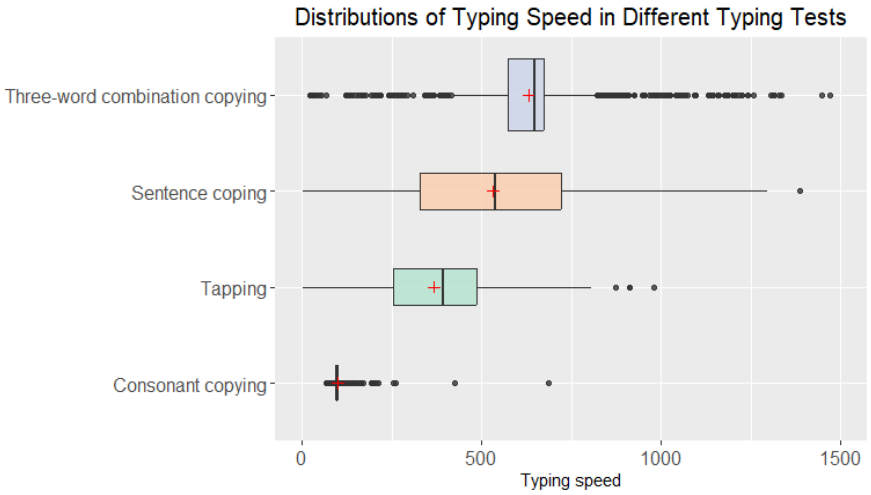

Figure 3. Box plots for typing speed scores in the typing tests.

3.2 Typing Tests Results

To showcase the writers' general typing skills, we calculated the average number of characters produced per minute in each typing test as a global typing speed measure. Figure 3 displays the distributions of the typing speed scores for each test in box plots, each displaying the mean, median, interquartile range (middle 50%), minimum, and maximum scores.

Specifically, the average typing scores for characters per minute are 367.50 (SD = 166.8), 531.25 (SD = 248.79), 631.5 (SD = 90), and 98.96 (SD = 95.62) for tapping, sentence copying, three-word combination copying, and consonant copying, respectively.

3.3 Argumentative Writing Keystrokes Descriptives

To give an overview of writers' keystroke activities as represented in the keystroke logs, we calculated each writer's total writing time during the argumentative writing task by measuring their active writing time, from the first keystroke to the last change made to the text before submission. We also counted the total numbers of keystrokes and mouse clicks they produced in the writing process. Furthermore, we calculated inter-keystroke intervals (IKIs, the gap time between two consecutive key presses; Chukharev-Hudilaninen et al., 2019) within the active writing time to showcase the general production rate of the writers. Figure 4 shows the distribution of these IKIs across the keystroke logs in the corpus. Most writers' IKIs concentrated between 200 and 1000 milliseconds. Table 4 presents the details of these statistics for the whole corpus and for

each prompt group. A series of Kruskal-Wallis tests revealed significant prompt effects on the number of keystrokes ($\chi^2(3) = 17.72$, $p < .001$) and IKIs ($\chi^2(3) = 13.04$, $p < .01$), but no significant effects on the total writing time or the number of mouse clicks. Specifically, writers tended to produce more strokes and shorter IKIs when writing about ***Appearance*** and ***Materialism*** compared to ***Happiness*** and ***Competition*** (See Appendix B for the details of these prompts).

Table 4. Descriptive statistics for keystroke logs in the whole corpus and for each prompt

		Mean	SD	Median	Min	Max	Range
Overall (N = 4992)	Total writing time (in minutes)	26.39	5.75	28.05	3.59	176.59	173
	Number of keystrokes	3364.73	1606.58	3021.5	421	18452	18031
	Number of mouse clicks	37.15	43.47	28	0	1328	1328
	IKIs	569.51	265.69	515.15	95.04	4296.62	4201.58
Appearance (n = 1234)	Total writing time (in minutes)	26.49	5.41	28.07	4.01	67.37	63.36
	Number of keystrokes	3446.72	1666	3096	1112	18452	17340
	Number of mouse clicks	36.62	54.08	27	1	1328	1327
	IKIs	554.02	251.67	508.66	95.04	3059.99	2964.95
Competition (n = 1136)	Total writing time (in minutes)	26.31	4.97	28.05	6.4	59.21	52.81
	Number of keystrokes	3286.79	1544.1	2910	729	11096	10367
	Number of mouse clicks	37.25	38.1	29	1	642	641
	IKIs	578.87	249.91	537.93	159.73	1928.18	1768.45
Happiness (n = 1329)	Total writing time (in minutes)	26.2	6.82	27.88	3.59	176.59	173

	Number of keystrokes	3280.82	1612.05	2920	421	11649	11228
	Number of mouse clicks	38.45	42.91	28	1	576	575
	IKIs	587.59	298.21	521.86	154.01	4296.62	4142.61
Materialism (*n* = 1293)	Total writing time (in minutes)	26.55	5.47	28.16	5.64	76.88	71.24
	Number of keystrokes	3441.19	1591.06	3114	635	13694	13059
	Number of mouse clicks	36.23	36.56	28	0	539	539
	IKIs	557.48	255.3	496.99	127.56	1929.59	1802.03

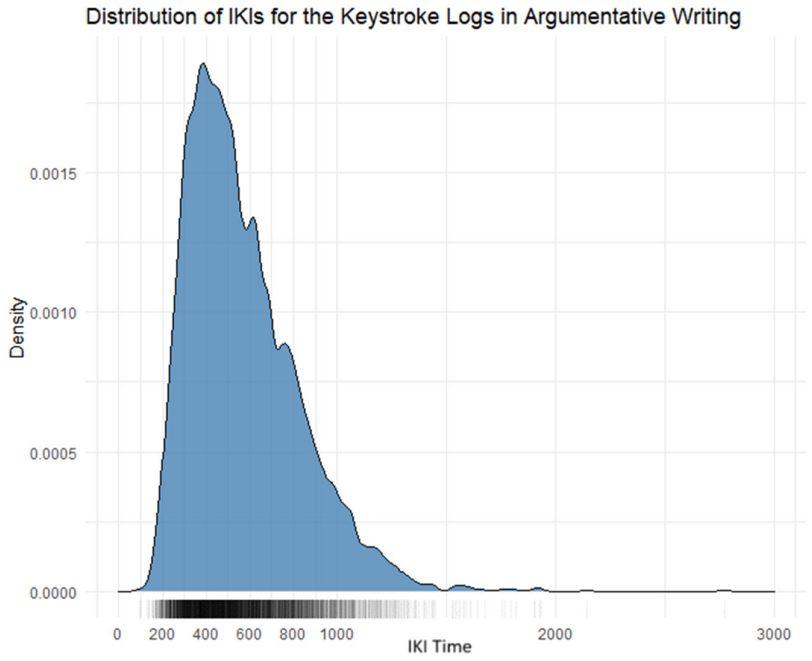

Figure 4. Writers' IKIs during argumentative writing

3.4 Holistic Writing Quality Scores for the Argumentative Texts

The distribution of the holistic writing quality scores is presented in Figure 5. As shown, the scores are generally normally distributed, ranging from 1 to 6. The average holistic writing score for the entire corpus is 3.72 (SD = 0.99).

Figure 6 shows the distribution of the holistic scores for different prompts. A one-way Analysis of Variance (ANOVA) was conducted to examine if there were score differences among different prompts. The ANOVA revealed a significant prompt effect on writing quality scores, $F(3, 4988) = 13.16$, $p < .001$. Post hoc tests revealed significant score differences between *Happiness* and *Materialism*, t = -5.45, $p < .001$, *Happiness* and *Appearance*, t = -5.15, $p < .000$, and *Happiness* and *Competition*, t = -4.45, $p < .001$. However, no significant score difference was found among *Materialism*, *Appearance*, and *Competition*.

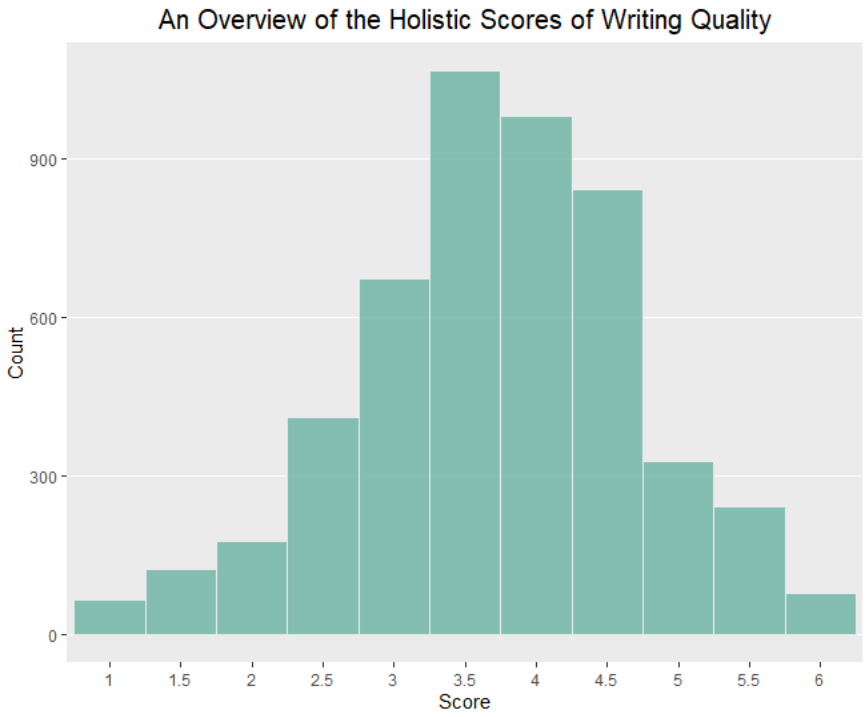

Figure 5. The distribution of holistic writing quality scores.

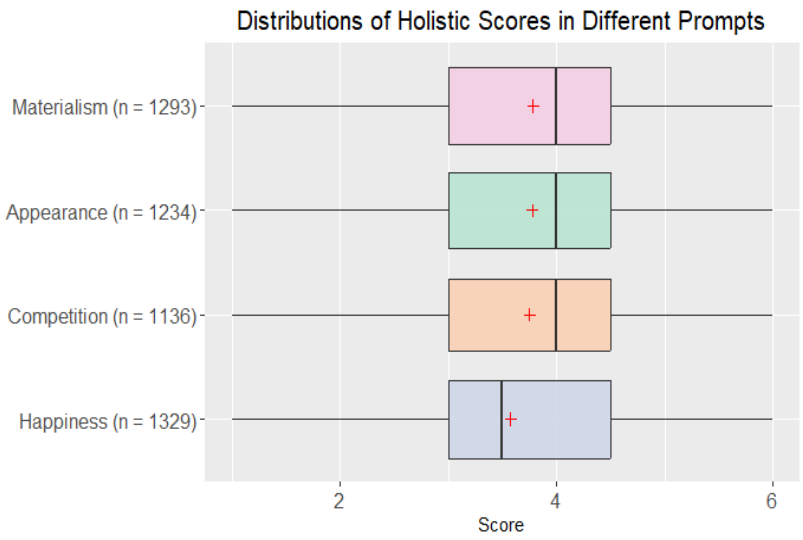

Figure 6. Box plots for score distribution in different prompts.

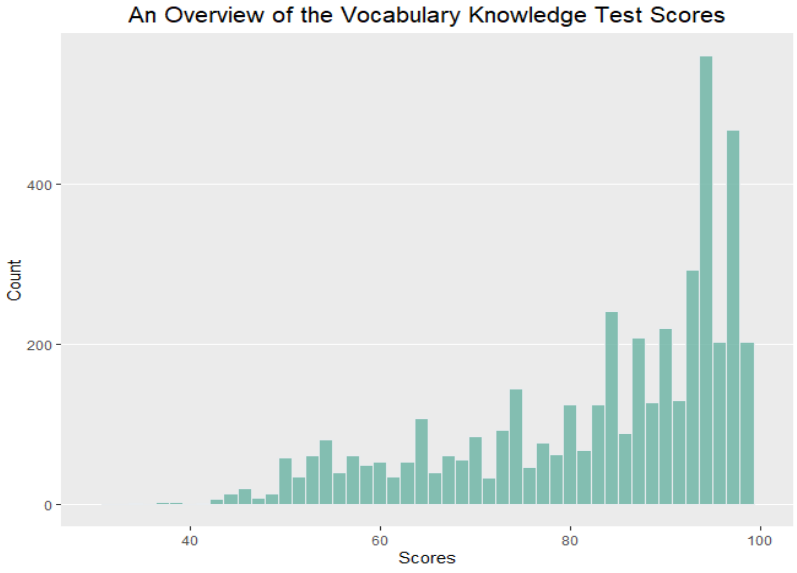

Figure 7. Distribution of vocabulary knowledge scores.

3.5 Vocabulary Knowledge Tests Results

To provide an overview of the writers' vocabulary knowledge as indicated in the LexTALE test, we calculated writers' LexTALE scores using a formula that corrects for the unequal proportion of words and nonwords: ((number of words correct / 40 * 100) + (number of nonwords correct / 20 * 100)) /2 (see www.lextale.com). The mean LexTALE score for the entire corpus is 84.96 (SD = 14.99) out of 100. The distribution of the LexTALE scores is shown in Figure 7.

4. An Illustrative Study Using a Subset of the KLiCKe Corpus

To validate the quality of the KLiCKe corpus and to illustrate how the corpus may be used for writing process research, we conducted a small-scale study to investigate the relationships between keystroke features and holistic writing quality for L1 and L2 writers, using a small subset of the entire corpus. We used regression analyses to assess the predictive strength of the keystroke features, typing test scores, and vocabulary knowledge scores in explaining writing quality scores.

4.1 Data Preparation and Keystroke Features Extraction

Four hundred data entries were sampled from the KLiCKe Corpus, which included argumentative texts and corresponding keystroke logs produced by both L1 (n = 200) and L2 (n = 200) writers, along with information about their typing skills and vocabulary knowledge.

To extract keystroke features from the logfiles, we first removed the inactive periods at the beginning (e.g., reading instructions, pre-planning) and the end of the processes (e.g., proof-reading, waiting for submission) to delimit the analyses to the actual text production and exclude noise that might skew the results of keystroke measures (e.g., long inactive periods might inflate mean pause lengths). We then analyzed the truncated logfiles (in .idfx formats) using Inputlog 9.0 to generate a set of keystroke indices in regard to writers' bursts, pausing behaviors, revision activities, and process variances as informed by previous research. To limit the inclusion of variables to those that are generalizable and facilitate the comparison of keystroke activities among writers, only keystroke measures based on means, proportions, or ratios were selected for statistical analyses. We also removed keystroke indices with null values (e.g., *mean length of between-sentence pauses*, *mean length of between-paragraph pauses*) for any participants in the KLiCKe corpus subset. As a result, a total of ten keystroke measures were retained. See Table 5 for descriptions of these ten measures.

We checked the keystroke measures prior to statistical analysis for any potential concerns. First, all measures were examined for the issue of near-zero variance since these variables are commonly considered to have little predictive power in regression models (Kuhn & Johnson, 2013).

Table 5. Descriptions of keystroke logging indices

Keystroke Logging Indices	Category	Descriptions
mean length of P-bursts	Burst	The mean length of the string of actions delineated by an initial and end pause exceeding 2 seconds and is measured in characters.
mean length of R-bursts	Burst	The mean length of the string of actions delineated by an insertion or deletion exceeding 2 seconds and is measured in characters.
mean length of pauses	Pause	The mean length of latencies between the previous and the current action that exceed 200ms in text production and is measured in seconds.
mean length of within-word pauses	Pause	The mean length of latencies within words that exceed 200ms in text production and is measured in seconds.
mean length of between-word pauses	Pause	The mean length of latencies between words that exceed 200ms in text production and is measured in seconds.
mean length of deletions	Revision	The mean number of characters deleted in one deletion activity.
proportion of deletions	Revision	The number of characters deleted divided by the total number of characters produced during the writing process.
mean length of insertions	Revision	The mean number of characters inserted in one insertion activity.
proportion of insertions	Revision	The number of characters inserted divided by the total number of characters produced during the writing process.
interval variance	Process variance	The standard deviation of the production rates for the 10 intervals in each writing process in relation to a task maximum.

To address this issue, the percentage of unique values and the frequency ratio of these unique values (i.e., the ratio of the frequency of the most prevalent value to that of the second most prevalent value) were calculated. None of the ten keystroke measures had a low percentage of unique values (< 10%) or a high frequency ratio (> 20). Second, a

series of correlation analyses were conducted among all the keystroke measures, writers' typing test and vocabulary knowledge test scores, and holistic scores of writing quality (see Appendix E for the correlation results). No measures were found to be highly collinear (i.e., absolute $r > .699$).

4.2 Data Analysis

To investigate whether keystroke measures were predictive of human rated holistic scores of writing quality for L1 and L2 writers, a series of linear regression models were built for the sampled dataset using *R* (R Core Team, 2020) and the CARET package (Kuhn et al., 2020). In these models, holistic scores for writing quality were entered as the dependent variable. The ten keystroke measures were entered as independent variables. The binary variable ***language*** (L1 vs. L2) was also included to control for potential effects of language nativeness on writing performance (e.g., Chenoweth & Hayes, 2001; Michel et al., 2020; Révész et al., 2022; Schoonen et al., 2003; Stevenson et al., 2006).Two-way interactions between ***language*** and each of the ten keystroke measures were also tested because the effect of keystroke measures might be moderated by writers' language nativeness. ***Typing speed*** and ***vocabulary knowledge*** were also entered as control variables to account for any effects they might exert on writing outcomes, as documented by previous studies (e.g., Barkoui, 2014; Berninger, 2000; Hayes & Berninger, 2014; Van Waes et al., 2019). In addition, the four-level categorical variable ***prompt*** (dummy coded) was also entered to control for prompt-based effects on writing quality.

To find the best fitting linear regression model, a backward variable elimination procedure was conducted manually following Hosmer, Lemeshow, and Sturdivant (2013). This had three steps: 1) a full model was built and a candidate predictor variable with the least significant *p* values in all sets of coefficients was selected. 2) The model was re-run without the candidate variable. If more than a 25% change was detected in the resulted coefficients, the variable was retained and the next non-significant variable was then chosen as a candidate variable. If no significant changes were found, the candidate variable was removed before proceeding to the next candidate variable. 3) This process was repeated until all non-significant variables were tested.

In order to better understand the contribution of each predictor as combined with other predictors in the final linear regression model, the relative importance of these predictors was assessed using the Lindeman, Merenda and Gold's (LMG) method (Lindeman et al., 1980) implemented in the R package "relaimpo" (Grυmping, 2006).

4.3 Results

Results from the final linear regression model are displayed in Table 6. The results show that writers' holistic scores were significantly predicted by ***typing speed***, ***vocabulary knowledge***, ***language***, and, to a lesser extent, by five keystroke measures: ***mean length of pauses***, ***mean length of between-word pauses***, ***proportion of deletions***, ***proportion of insertions***, and ***interval variance***. Specifically, the coefficients reported in the model

indicate that writers with higher essay scores tended to pause less in general and particularly between words. These writers were less likely to delete what they produced but were prone to engage in more insertions. In addition, their text production processes featured less variance in production fluency. The regression model also indicated that participants who produced higher-quality argumentative texts generally had better typing skills and vocabulary knowledge. Lastly, L1 writers scored significantly higher than L2 writers. However, no significant interaction was found between ***language*** and any of the keystroke measures, indicating that the effects of keystroke measures on holistic writing quality were similar across the two writer groups. The overall regression was statistically significant (adjusted $R^2 = 0.59$, $F(8, 391) = 72.67$, $p < .000$), indicating that the five keystroke measures along with ***language***, ***typing speed*** and ***vocabulary knowledge*** explained 59% percent of the variance in participants' holistic scores of writing quality.

Results of the relative importance of all predictors included in the final linear regression model are presented in Table 7. As shown, ***typing speed*** and ***vocabulary knowledge*** were the two most important variables in the model, contributing a total of 75.1% to the R^2 values. This was followed by ***language***, which accounted for 1.5% of the contribution on its own. The five keystroke logging measures in total made up 13.3% of the model. Among these keystroke measures, ***proportion of insertions*** and ***mean length of between-word pauses*** were identified as the two most important measures, followed by ***mean length of pauses*** and ***interval variance***. ***Proportion of deletions*** was the least important measure.

Table 6. Results of the final linear regression model

	coefficient	**standard error**	***T***	***p***
Intercept	1.506	0.253	5.946	0.000***
language: L2	-0.153	0.074	-2.085	0.038*
mean length of pauses	-0.227	0.047	-4.796	0.000***
mean length of between-word pauses	-0.083	0.029	-2.837	0.005**
proportion of deletions	-1.091	0.417	-2.619	0.009**
proportion of insertions	0.446	0.195	2.288	0.023*
interval variance	-1.592	0.535	-2.977	0.003**
typing speed	0.004	0.000	10.58	0.000***
vocabulary knowledge	0.024	0.003	9.379	0.000***

* $p < .05$, ** $p < .010$, *** $p < .001$

Table 7. Relative importance of predictors in the final regression model

Variables	Relative Importance
typing speed	37.8%
vocabulary knowledge	37.3%
language	11.5%
proportion of insertions	4.2%
mean length of between-word pauses	4.1%
mean length of pauses	3.3%
interval variance	1.2%
proportion of deletions	0.5%

4.4 Discussion

The illustrative study revealed that typing skills, vocabulary knowledge, and language status are among the most critical predictors of overall writing quality in the KLiCKe corpus. These findings align with prior research showing that higher writing quality is associated with stronger typing skills (e.g., Alvès et al., 2007; Barkaoui, 2014) and more advanced vocabulary knowledge (e.g., Albrechtsen et al., 2008; Milton et al., 2010). Furthermore, the results corroborate studies indicating that L1 writers tend to receive higher writing quality scores than their L2 counterparts (e.g., Crossley & McNamara, 2009; Ferris, 1994; Silva, 1993; Stevenson et al., 2006).

Additionally, a set of keystroke measures related to overall writing quality were identified. Specifically, our illustrative study indicated that higher writing scores in the KLiCKe corpus were associated with shorter pauses in general, shorter between-word pauses, lower proportion of deletions, higher proportion of insertions, and less process variance. These findings generally echo previous research on the relationship between writing process features and writing quality (e.g., Allen et al., 2016; Choi & Deane, 2020; Xu, 2018).

The results also provide evidence for the distinct roles of deleting and inserting behaviors in contributing to the final text quality and indicate that different cognitive operations might be involved when writers make deletions and insertions. It is generally acknowledged that revising is carried out at multiple levels and more experienced writers

tend to engage less in lower-level revising processes typified by local surface editing (e.g., Lindgren & Sullivan, 2006; Xu, 2018). A large proportion of deletions, often occurring at the point of inscription, may be related to low-level, convention- and rule-governed changes, such as spelling and grammar corrections. In contrast, insertions, typically made by moving the cursor to earlier text, are more likely related to changes at higher linguistic levels (word, clause, sentence) that alter meaning or manipulate content.

For our illustrative study, we used a set of generic keystroke measures calculated over the entire writing process, without segmenting features into specific time windows. However, using time window-based features—such as analyzing pauses or revision patterns at the beginning, middle, and end of the writing session—could offer a more nuanced understanding of how keystroke features relate to writing quality. Future studies could adopt fine-grained measures to capture distinct phases of the writing process, potentially revealing relationships not evident in our analysis, as demonstrated by Conijn et al. (in review).

5. Conclusion

This paper introduces the Keystroke Logging in Compositions for Knowledge Evaluation (KLiCKe) corpus, a large-scale dataset featuring detailed keystroke logs from ~5,000 argumentative essays written by adult English writers in the United States. The corpus records each keystroke and mouse operation, along with corresponding time stamps and cursor position information, using a web-based keystroke logging program. Data are provided in .csv format (for custom analysis) and .idfx format (compatible with Inputlog). Holistic writing quality scores for all essays are included, derived from double-blind ratings by trained human raters using a standardized grading scale. Additionally, the corpus offers demographic details on the writers, including age, gender, native language, ethnicity/race, education level, as well as typing skills and vocabulary knowledge. As a publicly available resource, KLiCKe bridges gaps in process-oriented writing research and offers new possibilities for advancing writing assessment and instruction.

Using the KLiCKe corpus, researchers can develop fine-grained measures that combine rich linguistic features (e.g., part-of-speech tagging, word frequency metrics, syntactic complexity indices) with temporal features such as pause durations. These advanced measures could improve the granularity of writing process analyses, providing deeper insights into the cognitive processes involved in text production. Additionally, the large scale and detailed nature of the KLiCKe corpus make it an ideal resource for training machine learning algorithms to identify latent patterns in the writing process that may be difficult to detect through traditional methods. Researchers could use this dataset to develop predictive models linking writing dynamics to writing quality, classify writing behaviors, or investigate individual differences in writing strategies.

Additionally, KLiCKe could accelerate the integration of keystroke-logged information into AWE tools and other writing platforms, enabling process-based evaluations and real-time feedback tailored to learners' needs during text production. For

example, a writing platform could use keystroke logs to identify indicators of writing challenges, such as prolonged pauses (which may signal mind-wandering or difficulties in idea generation) or frequent revisions (which might suggest struggles with idea formulation or spelling, depending on the revision type). This information could facilitate interventions to support writers during challenging stages of the writing process. Furthermore, KLiCKe could inspire more process-oriented instructional strategies. By incorporating keystroke logs into diagnostic techniques, teachers could monitor learners' writing development more effectively. The insights gained from analyzing keystroke logs could provide valuable evidence of learners' writing strategies and difficulties, thus guiding instructional decisions.

However, the corpus is not without limitations. First, in its design, KLiCKe relies on a timed independent argumentative writing task, which offers a window into writers' text production. While this writing task resembles writing in well-known standardized assessment settings such as SAT, TOFEL, and IELTS, it does not fully capture the range of writing practices, particularly those encountered in personal, professional, and academic contexts. In these settings, the writing process features more recursive characteristics with no strict time limits for planning, drafting, and revising activities. Additionally, the corpus does not account for collaborative writing and source-based writing, both of which are key context of academic writing (Ferretti & Lewis, 2018). Another limitation is the significant prompt effect observed on some general process features (e.g., number of keystrokes, IKIs) as well as the holistic writing quality. A possible contributing factor could be the varying levels of topic knowledge among writers for different prompts, which may influence both text production and writing outcomes (e.g., He & Shi, 2012; Liu & Stapleton, 2018; Yang & Kim, 2020; Yoon, 2021). For example, it is likely that KLiCKe essay writers had varying levels of understanding or insight into topics like *Happiness* compared to other topics. As such, researchers using the KLiCKe corpus for writing quality studies should consider potential prompt effects. Finally, despite efforts to minimize uncontrolled variables through design and instruction, we acknowledge the inherent limitations of crowd-sourced data collection, such as environmental factors like noise levels, visual disturbances, and hardware configurations, which may have influenced participants' writing processes.

Despite its limitations, we envision the KLiCKe corpus as a significant advancement for writing research. In our illustrative study, we used a small subset of KLiCKe to investigate the relationships between keystroke measures of the writing process and the holistic writing quality of argumentative texts produced by both L1 and L2 writers. However, the potential of the KLiCKe corpus extends far beyond this initial exploration. First and foremost, the keystroke logs in KLiCKe can be utilized to model various writer behaviors during text production, such as writing strategies, challenges, and levels of engagement. Additionally, the dynamics of the writing process can be mapped to the production of a wide range of textual features, offering deeper insights into the connections between writing behaviors and writing outcomes. This includes, but is not

limited to, specific linguistic units or characteristics (e.g., formulaic expressions, organizational markers, and summary statements), the relational structures of discourse elements (e.g., how a claim supported by evidence may reinforce higher-level arguments), and text cohesion. Moreover, the rich metadata available in KLiCKe—such as writers' typing skills, vocabulary knowledge, and demographic information—can be incorporated to investigate how various socio-cognitive factors moderate the relationships between the writing process and writing outcomes. Such an approach holds promise for furthering our understanding of writing as a complex cognitive and social activity.

Acknowledgements

The dataset generated and analyzed during this study, including the KLiCKe corpus, is publicly available for researchers interested in further exploration. The corpus is hosted on GitHub and can be accessed at https://github.com/terryyutian/KLiCKe-Corpus. For additional details regarding its construction and applications, please contact Yu Tian at ytian126@asu.edu. We encourage researchers to utilize this dataset to advance the understanding of writing processes and contribute to the broader body of knowledge in this field.

This work was supported by funding from Schmidt Futures, Duolingo English Test, and the Adult Literacy Research Center at Georgia State University. The authors declare no conflicts of interest pertaining to this study.

References

Albrechtsen, D., Haastrup, K., & Henriksen, B. (2008). *Vocabulary and writing in a first and Second language: Processes and development*. Palgrave Macmillan. http://dx.doi.org/10.1057/9780230593404

Allen, L. K., Jacovina, M. E., Dascalu, M., Roscoe, R. D., Kent, K. M., Likens, A. D., & McNamara, D. S. (2016). {ENTER} ing the Time Series {SPACE}: Uncovering the Writing Process through Keystroke Analyses. *International Educational Data Mining Society*.

Almond, R., Deane, P., Quinlan, T., Wagner, M., & Sydorenko, T. (2012). A preliminary analysis of keystroke log data from a timed writing task. *ETS Research Report Series*, *2012*(2), i-61. http://dx.doi.org/10.1002/j.2333-8504.2012.tb02305.x

Alvès, R.A., Castro, S.L., & de Sousa, L. (2007). Influence of typing skill on pause–execution cycles in written composition. In Rijlaarsdam, G. (Series Ed.); M. Torrance, L. van Waes, & D. Galbraith (Volume Eds.), *Writing and Cognition: Research and Applications* (Studies in Writing, Vol. 20, pp. 55–65). Amsterdam: Elsevier. http://dx.doi.org/10.1163/9781849508223_005

Barkaoui, K. (2014). Examining the impact of L2 proficiency and keyboarding skills on scores on TOEFL-iBT writing tasks. *Language Testing*, *31*(2), 241-259. http://dx.doi.org/10.1177/0265532213509810

Barkaoui, K. (2016). What and when second-language learners revise when responding to timed writing tasks on the computer: The roles of task type, second language proficiency, and keyboarding skills. *The Modern Language Journal*, *100*(1), 320-340. http://dx.doi.org/10.1111/modl.12316

Barkaoui, K. (2019). What can l2 writers'pausing behavior tell us about their l2 writing processes? *Studies in Second Language Acquisition*, *41*(3), 529-554. http://dx.doi.org/10.1017/s027226311900010x

Berninger, V. (2000). Development of language by hand and its connections to language by ear, mouth, and eye. *Topics of Language Disorders*, *20*, 65–84. http://dx.doi.org/10.1097/00011363-200020040-00007

Bowen, N. E. J. A., Thomas, N., & Vandermeulen, N. (2022). Exploring feedback and regulation in online writing classes with keystroke logging. *Computers and Composition*, *63*, 102692. http://dx.doi.org/10.1016/j.compcom.2022.102692

Caporossi, G., Leblay, C., & Usoof, H. (2023) GenoGraphiX-LOG (Version 2.1.0) [Computer software]. HEC Montréal & University of Turku. https://ggxlog.net

Carl, M (2012). Translog-II: a program for recording user activity data for empirical reading and writing research. In *Proceedings of the eighth international conference on language resources and evaluation (LREC12)*, pp 4108–4112.

Chenoweth, N. A., & Hayes, J. R. (2001). Fluency in writing: Generating text in L1 and L2. *Written Communication*, 18, 80-98. http://dx.doi.org/10.1177/0741088301018001004

Choi, I., & Deane, P. (2020). Evaluating Writing Process Features in an Adult EFL Writing Assessment Context: A Keystroke Logging Study. *Language Assessment Quarterly*, 1-26. http://dx.doi.org/10.1080/15434303.2020.1804913

Chukharev-Hudilainen, E. (2019). Empowering automated writing evaluation with keystroke logging. In E. Lindgren & K. P. H. Sullivan (Eds.), *Observing writing* (pp.125-142). Brill. http://dx.doi.org/10.1163/9789004392526_007

Chukharev-Hudilainen, E., Saricaoglu, A., Torrance, M., & Feng, H. H. (2019). Combined deployable keystroke logging and eyetracking for investigating L2 writing fluency. *Studies in Second Language Acquisition*, 41(3), 583-604. http://dx.doi.org/10.1017/s027226311900007x

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1), 37-46. http://dx.doi.org/10.1177/001316446002000104

Conijn, R., Cook, C., Van Zaanen, M., & Van Waes, L. (2022). Early prediction of writing quality using keystroke logging. *International Journal of Artificial Intelligence in Education*, *32*(4), 835-866. http://dx.doi.org/10.1007/s40593-021-00268-w

Conijn, R., Rossetti, A., Vandermeulen, N., & Van Waes, L. (n.d.). *Phase to phase: Towards an automated procedure to identify phases in writing processes using keystroke data*. SSRN. https://ssrn.com/abstract=4993558 or https://doi.org/10.2139/ssrn.4993558

Crossley, S. A., & McNamara, D. S. (2009). Computational assessment of lexical differences in L1 and L2 writing. *Journal of second language writing*, *18*(2), 119-135. http://dx.doi.org/10.1016/j.jslw.2009.02.002

Crossley, S., Tian, Y., & Wan, Q. (2022). Argumentation features and essay quality: Exploring relationships and incidence counts. *Journal of Writing Research*, *14*(1), 1–34. http://dx.doi.org/10.17239/jowr-2022.14.01.01

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. http://dx.doi.org/10.1002/ets2.12002

Deane, P., & Quinlan, T. (2010). What automated analyses of corpora can tell us about students' writing skills. *Journal of Writing Research*, *2*(2), 151-177. http://dx.doi.org/10.17239/jowr-2010.02.02.4

Deane, P., & Zhang, M. (2015). Exploring the feasibility of using writing process features to assess text production skills. *ETS Research Report Series*, *2015*(2), 1-16. http://dx.doi.org/10.1002/ets2.12071

Dragsted, B., & Carl, M. (2013). Towards a classification of translation styles based on eye-tracking and keylogging data. *Journal of Writing Research*, *5*(1). http://dx.doi.org/10.17239/jowr-2013.05.01.6

Ferretti, R. P., & Lewis, W. E. (2018). Argumentative writing. In S. Graham, C. A. MacArthur, & J. Fitzgerald (Eds.) *Best practices in writing instruction* (pp.135-162). Guilford. http://dx.doi.org/10.17239/jowr-2014.06.02.5

Ferris, D. R. (1994). Rhetorical strategies in student persuasive writing: Differences between native and non-native English speakers. *Research in the Teaching of English*, 45-65. http://dx.doi.org/10.58680/rte199415388

Frid, J., Wengelin, A., Johansson, V., Johansson, R., & Johansson, M. (2012, July). Testing the temporal accuracy of keystroke logging using the sound card. Paper presented at the 13th International EARLI SIG Writing Conference, Porto, Portugal.

Galbraith, D., & Baaijen, V. M. (2019). Aligning keystrokes with cognitive processes in writing. In E. Lindgren & K. P. H. Sullivan (Eds.), *Observing Writing* (pp. 306-325). Brill. http://dx.doi.org/10.1163/9789004392526_015

Grammarly. (2022). *Grammarly [English writing assistant software]*. San Francisco, CA: Grammarly Inc. https://www.grammarly.com

Grumping, U. (2006). R package relaimpo: relative importance for linear regression. *J. Stat. Softw*, *17*(1), 139-147. http://dx.doi.org/10.18637/jss.v017.i01

Guo, H., Zhang, M., Deane, P., & Bennett, R. E. (2019). Writing process differences in subgroups reflected in keystroke logs. *Journal of Educational and Behavioral Statistics*, *44*(5), 571-596. http://dx.doi.org/10.3102/1076998619856590

Hamel, M. J., & Séror, J. (2016). Video screen capture to document and scaffold the L2 writing process. *Language-learner computer interactions: Theory, methodology, and applications*, 137-162. http://dx.doi.org/10.1075/lsse.2.07ham

Hayes, J. R., & Flower, L. (1981). *Uncovering cognitive processes in writing: An introduction to protocol analysis*. ERIC Clearinghouse.

Hayes, J. R., & Berninger, V. W. (2014). Cognitive processes in writing: A framework. In B. Arfe, J., Dockrell, & V. W. Berninger (Eds), *Writing development in children with hearing loss, dyslexia, or oral language problems* (pp. 3–15). Oxford: Oxford University Press. http://dx.doi.org/10.1093/acprof:oso/9780199827282.001.0001

He, L., & Shi, L. (2012). Topical knowledge in ESL writing. *Language Testing*, 29, 443–464. http://dx.doi.org/10.1177/0265532212436659

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (Vol. 398). John Wiley & Sons. http://dx.doi.org/10.32614/cran.package.aplore3

Janssen, D., Van Waes, L., & Van den Bergh, H. (2013). Effects of thinking aloud on writing processes. In *The science of writing* (pp. 233-250). Routledge.

Jung, J. (2017). Effects of task complexity on L2 writing processes and linguistic complexity: A keystroke logging study. *English Teaching*, *72*(4), 179-200. http://dx.doi.org/10.15858/engtea.72.4.201712.179

Kaufer, D. S., Hayes, J. R., & Flower, L. (1986). Composing written sentences. *Research in the Teaching of English*, 20, 121-140. http://dx.doi.org/10.58680/rte198615612

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*, New York, NY: Springer.

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., ... & Team, R. C. (2020). Package 'caret'. *The R Journal*, *223*(7).

Leijten, M., & Van Waes, L. (2006). Inputlog: New perspectives on the logging of on-line writing processes in a Windows environment. In *Computer key-stroke logging and writing* (pp. 73-93). Brill. http://dx.doi.org/10.1163/9780080460932_006

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. http://dx.doi.org/10.1177/0741088313491692

Leijten, M., & Van Waes, L. (2020). Designing keystroke logging research in writing studies. *Chinese journal of second language writing*, *1*(1), 18-39. http://dx.doi.org/10.1163/9789004392526_005

Leijten, M., Van Waes, L., Schriver, K., & Hayes, J. R. (2013). Writing in the workplace: Constructing documents using multiple digital sources. *Journal of Writing Research*, *5*(3). http://dx.doi.org/10.17239/jowr-2014.05.03.3

Lemhufer, K., & Broersma, M. (2012). Introducing LexTALE: A quick and valid lexical test for advanced learners of English. *Behavior research methods*, *44*, 325-343. http://dx.doi.org/10.3758/s13428-011-0146-0

Lindeman, R. H., Merenda, P. F., & Gold, R. Z. (1980). *Introduction to bivariate and multivariate analysis*. Scott Foresman.

Lindgren, E., & Sullivan, K. P. H. (2003). Stimulated recall as a trigger for increasing noticing and language awareness in the L2 writing classroom: A case study of two young female writers. *Language Awareness*, 12(3–4), 172–186. http://dx.doi.org/10.1080/09658410308667075

Lindgren, E. & Sullivan K. P. H. (2006). Analyzing on-line revision. In G. Rijlaarsdam (Series Ed.) and K. P. H. Sullivan, & E. Lindgren. (Vol. Eds.), Studies in Writing, Vol.18, *Computer Keystroke Logging: Methods and Applications*, (157–188). Oxford: Elsevier. http://dx.doi.org/10.1163/9780080460932_010

Lindgren, E., Sullivan, K. P. H., & Stevenson, M. (2008). Supporting the reflective language learner with computer keystroke logging. In B. Barber & F. Zhang (Eds.), *Handbook of research on computer enhanced language acquisition and learning* (pp. 189–204). Hershey, NY: Information Science Reference, IGI Global. http://dx.doi.org/10.4018/978-1-59904-895-6.ch011

Lindgren, E., & Sullivan, K. (Eds.). (2019). *Observing writing: Insights from keystroke logging and handwriting*. Leiden, The Netherlands: Brill. http://dx.doi.org/10.1163/9789004392526

Liu, F., & Stapleton, P. (2014). Counterargumentation and the cultivation of critical thinking in argumentative writing: Investigating washback from a high-stakes test. *System*, 45, 117–128. http://dx.doi.org/10.1016/j.system.2014.05.005

Mackey, A., & Gass, S. M. (2015). *Second language research: Methodology and design*. Routledge.

Medimorec, S., & Risko, E. F. (2016). Effects of disfluency in writing. *British Journal of Psychology*, *107*(4), 625-650. http://dx.doi.org/10.1111/bjop.12177

Michel, M., Révész, A., Lu, X., Kourtali, N. E., Lee, M., & Borges, L. (2020). Investigating L2 writing processes across independent and integrated tasks: A mixed-methods study. *Second Language Research*, *36*(3), 307-334. http://dx.doi.org/10.1177/0267658320915501

Miletić, A., Benzitoun, C., Cislaru, G., & Herrera-Yanez, S. (2022, June). Pro-text: An annotated corpus of keystroke logs. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 1732-1739).

Milton, J., Wade, J., & Hopkins, N. (2010). Aural word recognition and oral competence in English as a foreign language. *Insights into non-native vocabulary teaching and learning*, *52*, 83-98. http://dx.doi.org/10.21832/9781847692900-007

Mucoz Martín, R., & Apfelthaler, M. (2022). A Task Segment Framework to study keylogged translation processes. *Translation & Interpreting*, *14*(2), 8-31. http://dx.doi.org/10.12807/ti.114202.2022.a02

Olive, T., Favart, M., Beauvais, C., & Beauvais, L. (2009). Children's cognitive effort and fluency in writing: Effects of genre and of handwriting automatisation. *Learning and Instruction*, *19*(4), 299-308. http://dx.doi.org/10.1016/j.learninstruc.2008.05.005

Qualtrics. (2022). *Qualtrics [Online survey platform]*. Provo, UT: Qualtrics. https://www.qualtrics.com

R Core Team (2020). R: a language and environment for statistical computing. R *Foundation for Statistical Computing*. Retrieved from https://www.r-project.org/.

Ranalli, Jim, Feng, Hui-Hsien, & Chukharev-Hudilainen, Evgeny. (2018). Exploring the potential of process-tracing technologies to support assessment for learning of L2 writing. *Assessing Writing*, 36, 77–89. http://dx.doi.org/10.1016/j.asw.2018.03.007

Ransdell, S., Arecco, M. R., & Levy, C. M. (2001). Bilingual long-term working memory: The effects of working memory loads on writing quality and fluency. *Applied Psycholinguistics*, *22*(1), 113. http://dx.doi.org/10.1017/s0142716401001060

Révész, A., Michel, M., & Lee, M. (2017). *Investigating IELTS academic writing task 2: Relationship between cognitive writing processes, text quality, and working memory*. IELTS Research Reports Online Series. https://www.ielts.org/en-us/for-researchers/research-reports/ielts_online_rr_2017-3

Révész, A., Michel M., & Lee, M. (2022). Exploring the relationship of working memory to the temporal distribution of pausing and revision behaviors during L2 writing. *Studies in Second Language Acquisition*. 45(3), 680-709. http://dx.doi.org/10.1017/s0272263123000074

Russo, J. E., Johnson, E. J., & Stephens, D. L. (1989). The validity of verbal protocols. *Memory & cognition*, *17*(6), 759-769. http://dx.doi.org/10.3758/bf03202637

Schoonen, R., Gelderen, A. V., Glopper, K. D., Hulstijn, J., Simis, A., Snellings, P., & Stevenson, M. (2003). First language and second language writing: The role of linguistic knowledge, speed of processing, and metacognitive knowledge. *Language learning*, *53*(1), 165-202. http://dx.doi.org/10.1111/1467-9922.00213

Schumacher, G. M., Klare, G. R., Cronin, F. C., & Moses, J. D. (1984). Cognitive activities of beginning and advanced college writers: A pausal analysis. *Research in the Teaching of English*, 169-187. http://dx.doi.org/10.58680/rte198415678

Silva, T. (1993). Toward an understanding of the distinct nature of L2 writing: The ESL research and its implications. *TESOL Quarterly*, 27, 657–675. http://dx.doi.org/10.2307/3587400

Sinharay, S., Zhang, M., & Deane, P. (2019). Prediction of essay scores from writing process and product features using data mining methods. *Applied Measurement in Education*, *32*(2), 116-137. http://dx.doi.org/10.1080/08957347.2019.1577245

Stannard, R. (2019). A review of screen capture technology feedback research. *Studia Universitatis Babes-Bolyai-Philologia*, *64*(2), 61-72. http://dx.doi.org/10.24193/subbphilo.2019.2.05

Stevenson, M., Schoonen, R., & De Glopper, K. (2006). Revising in two languages: A multi-dimensional comparison of online writing revisions in L1 and FL. *Journal of Second Language Writing*, 15(3), 201-233. http://dx.doi.org/10.1016/j.jslw.2006.06.002

Strumqvist, S., Holmqvist, K., Johansson, V., Karlsson, H., & Wengelin, A. (2006). What key-logging can reveal about writing. In K. P. H. Sullivan & E. Lindgren (Eds.), *Computer keystroke logging and writing: Methods and applications* (pp. 45-72). Amsterdam, Netherlands: Elsevier. http://dx.doi.org/10.1163/9780080460932_005

Strumqvist, S., & Malmsten, L. (1998). ScriptLog Pro 1.04: User's manual. *Technical. Göteborg: University of Göteborg*.

Tian, Y., Kim, M., & Crossley, S. (2024). Making sense of L2 written argumentation with keystroke logging. *Journal of Writing Research*, *15*(3), 435-461. http://dx.doi.org/10.17239/jowr-2024.15.03.01

Tian, Y., Kim, M., Crossley, S., & Wan, Q. (2021). Cohesive devices as an indicator of L2 students' writing fluency. *Reading and Writing*, 1-23. http://dx.doi.org/10.1007/s11145-021-10229-3

Van den Bergh, H., & Rijlaarsdam, G. (2007). The dynamics of idea generation during writing: An online study. In *Writing and cognition* (pp. 125-150). Brill. http://dx.doi.org/10.1163/9781849508223_010

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. http://dx.doi.org/10.1016/j.compcom.2015.09.012

Van Waes, L., Leijten, M., Pauwaert, T., & Van Horenbeeck, E. (2019). A multilingual copy task: Measuring typing and motor skills in writing with Inputlog. *Journal of open research software.-2013, currens*, *7*(30), 1-8. http://dx.doi.org/10.5334/jors.234

Van Waes, L., Leijten, M., Van Horenbeeck, E., & Pauwaert, T. (2012). A generic XML-structure for logging human computer interaction. In *13th International EARLI SIG Writing Conference, Porto, Portugal*.

Vandermeulen, N., Leijten, M., & Van Waes, L. (2020). Reporting writing process feedback in the classroom: Using keystroke logging data to reflect on writing processes. *Journal of Writing Research*, *12*(1), 109-140. http://dx.doi.org/10.17239/jowr-2020.12.01.05

Vandermeulen, N., Van Steendam, E., De Maeyer, S., & Rijlaarsdam, G. (2023). Writing process feedback based on keystroke logging and comparison with exemplars: Effects on the quality and process of synthesis texts. *Written Communication*, *40*(1), 90-144. http://dx.doi.org/10.1177/07410883221127998

Vandermeulen, N., Van Steendam, & E., Rijlaarsdam, G. (2020). DATASET - Baseline data LIFT Synthesis Writing project [Data set]. Zenodo. https://doi.org/10.5281/zenodo.3893538

Wengelin, E. (2006). Examining pauses in writing: Theories, methods and empirical data. In K.P.H. Sullivan & E. Lindgren (Eds.), *Computer Key-Stroke Logging and Writing: Methods and Applications* (pp. 107-130). Amsterdam, the Netherlands: Elsevier. http://dx.doi.org/10.1163/9780080460932_008

Wengelin, E., Frid, J., Johansson, R., & Johansson, V. (2019). Combining keystroke logging with other methods. Towards an experimental environment for writing process research. In E. Lindgren & K. Sullivan (Eds.), *Observing writing: Insights from keystroke logging and handwriting* (pp. 30–49). Leiden: Brill. http://dx.doi.org/10.1163/9789004392526_003

Wengelin, E., Johansson, R., Frid, J., & Johansson, V. (2024). Capturing writers' typing while visually attending the emerging text: A methodological approach. *Reading and Writing*, *37*(2), 265-289. http://dx.doi.org/10.1007/s11145-022-10397-w

Xu, C. (2018). Understanding online revisions in L2 writing: A computer keystroke-log perspective. *System*, *78*, 104-114. http://dx.doi.org/10.1016/j.system.2018.08.007

Yang, W., & Kim, Y. (2020). The effect of topic familiarity on the complexity, accuracy, and fluency of second language writing. *Applied Linguistics Review*, 11, 79–108. http://dx.doi.org/10.1515/applirev-2017-0017

Yoon, H. J. (2021). Interactions in EFL argumentative writing: Effects of topic, L1 background, and L2 proficiency on interactional metadiscourse. *Reading and Writing*, *34*(3), 705-725. http://dx.doi.org/10.1007/s11145-020-10085-7

Zhang, M., Hao, J., Li, C., & Deane, P. (2016). Classification of writing patterns using keystroke logs. In *Quantitative psychology research: The 80th annual meeting of the psychometric society, Beijing, 2015* (pp. 299-314). Springer International Publishing. http://dx.doi.org/10.1007/978-3-319-38759-8_23

Zhu, M., Zhang, M., & Deane, P. (2019). Analysis of keystroke sequences in writing logs. *ETS Research Report Series*, *2019*(1), 1-16. http://dx.doi.org/10.1002/ets2.12247

Appendix A: Title: Demographic Survey Questions

Session One: Questions for both L1 and L2 participants

1. What is your Amazon Mechanical Turk worker ID?
2. What gender do you identify as?
3. What is your age?
4. What is your country citizenship?
5. Which race / ethnicity best describes you?
6. What is the highest degree or level of school you have completed (If currently enrolled, highest degree received)?
7. Are you right-handed/left-handed/ambidexterous?
8. Which of the following applies to you?

A. I grew up speaking English / I am a native English speaker.
B. I grew up speaking language(s) other than English / I am a non-native speaker of English

Session Two: Questions for L2 participants only (If option B is selected for question 8)

9. What is your native language or mother tongue (should not be English)?
10. At what age did you start learning English? (Please write the number only. e.g., 6).
11. How many years have you studied English? (Please write the number only. e.g., 8.5).
12. How many years have you been in the U.S.? (Please write the number only. e.g., 2).

Appendix B: SAT-based Writing Prompts Used in the Study

Topic	Writing Prompt
Appearance	All around us appearances are mistaken for reality. Clever advertisements create favorable impressions but say little or nothing about the products they promote. In stores, colorful packages are often better than their contents. In the media, how certain entertainers, politicians, and other public figures appear is more important than their abilities. All too often, what we think we see becomes far more important than what really is. Do images and impressions have too much of an effect on people?
Competition	While some people promote competition as the only way to achieve success, others emphasize the power of cooperation. Intense rivalry at work or play or engaging in competition involving ideas or skills may indeed drive people either to avoid failure or to achieve important victories. In a complex world, however, cooperation is much more likely to produce significant, lasting accomplishments. Do people achieve more success by cooperation or by competition?
Happiness	Some believe that happiness comes by pursuing their dreams and their own personal goals. Others believe that people are happy only when they have their minds fixed on some goal other than their own happiness. Accordingly, happiness comes when people focus on the happiness of others or on the improvement of humanity. Aiming at something other than their own happiness, they find happiness along the way. Are people more likely to be happy if they focus on their personal goals or on the happiness of others?
Materialism	Materialism: it's the thing that everybody loves to hate. Few aspects of modern life have been more criticized than materialism. But let's face it: materialism—acquiring possessions and spending money—is a vital source of meaning and happiness in our time. People may criticize modern society for being too materialistic, but the fact remains that most of us spend most of our energy producing and consuming more and more stuff. Should modern society be criticized for being materialistic?

Appendix C: Evaluating the Temporal Accuracy of Our Web-based Keystroke Logging Program

To evaluate the temporal accuracy of the keystroke logging program developed for constructing the KLiCKe corpus, we followed the research protocol outlined by Frid et al. (2012). The assessment involved two trials, each requiring the Space bar to be pressed 50 times. We recorded the timestamps of these keypresses using our custom program under two different web browsers: Google Chrome and Mozilla Firefox. At the same time, we captured the auditory signal of the keypresses using a standard laptop sound card.

The recorded audio was analyzed by manually annotating the waveform to pinpoint the exact timestamps corresponding to each keypress. To assess the timing accuracy, we computed both point-by-point differences and interval differences between the timestamps generated by our program and those obtained from the audio annotations.

For validation, we replicated this procedure using Inputlog 9. We present the results in the table below.

Table B1. Temporal accuracy test results in seconds

	Point-by-point differences			Interval differences		
	SD	range	max	SD	range	max
Our program + Chrome	0.001	0.006	0.006	0.001	0.006	0.006
Our program + Firefox	0.001	0.003	0.003	0.002	0.006	0.006
Inputlog	0.001	0.006	0.006	0.001	0.003	0.003

Note. SD = standard deviation; range = the range of differences; max = maximum absolute difference.

Appendix D: Holistic Rating Form

After reading each essay and completing the analytical rating form, assign a holistic score based on the rubric below. For the following evaluations you will need to use a grading scale between 1 (minimum) and 6 (maximum). As with the analytical rating form, the distance between each grade (e.g., 1-2, 3-4, 4-5) should be considered equal.

SCORE OF 6: An essay in this category demonstrates clear and consistent mastery, although it may have a few minor errors. A typical essay effectively and insightfully develops a point of view on the issue and demonstrates outstanding critical thinking, using clearly appropriate examples, reasons, and other evidence to support its position; is well organized and clearly focused, demonstrating clear coherence and smooth progression of ideas; exhibits skillful use of language, using a varied, accurate, and apt vocabulary; demonstrates meaningful variety in sentence structure; is free of most errors in grammar, usage, and mechanics.

SCORE OF 5: An essay in this category demonstrates reasonably consistent mastery, although it will have occasional errors or lapses in quality. A typical essay effectively develops a point of view on the issue and demonstrates strong critical thinking, generally using appropriate examples, reasons, and other evidence to support its position; is well organized and focused, demonstrating coherence and progression of ideas; exhibits facility in the use of language, using appropriate vocabulary; demonstrates variety in sentence structure; is generally free of most errors in grammar, usage, and mechanics.

SCORE OF 4: An essay in this category demonstrates adequate mastery, although it will have lapses in quality. A typical essay develops a point of view on the issue and demonstrates competent critical thinking, using adequate examples, reasons, and other evidence to support its position; is generally organized and focused, demonstrating some coherence and progression of ideas; exhibits adequate but inconsistent facility in the use of language, using generally appropriate vocabulary; demonstrates some variety in sentence structure; has some errors in grammar, usage, and mechanics.

SCORE OF 3: An essay in this category demonstrates developing mastery, and is marked by ONE OR MORE of the following weaknesses: develops a point of view on the issue, demonstrating some critical thinking, but may do so inconsistently or use inadequate examples, reasons, or other evidence to support its position; is limited in its organization or focus, or may demonstrate some lapses in coherence or progression of ideas; displays developing facility in the use of language, but sometimes uses weak vocabulary or inappropriate word choice; lacks variety or demonstrates problems in sentence structure; contains an accumulation of errors in grammar, usage, and mechanics.

SCORE OF 2: An essay in this category demonstrates little mastery, and is flawed by ONE OR MORE of the following weaknesses: develops a point of view on the issue that is vague or seriously limited, and demonstrates weak critical thinking, providing inappropriate or insufficient examples, reasons, or other evidence to support its position; is poorly organized and/or focused, or demonstrates serious problems with coherence or progression of ideas; displays very little facility in the use of language, using very limited vocabulary or incorrect word choice; demonstrates frequent problems in sentence structure; contains errors in grammar, usage, and mechanics so serious that meaning is somewhat obscured.

SCORE OF 1: An essay in this category demonstrates very little or no mastery, and is severely flawed by ONE OR MORE of the following weaknesses: develops no viable point of view on the issue, or provides little or no evidence to support its position; is disorganized or unfocused, resulting in a disjointed or incoherent essay; displays fundamental errors in vocabulary; demonstrates severe flaws in sentence structure; contains pervasive errors in grammar, usage, or mechanics that persistently interfere with meaning.

Holistic score based on attached rubric (1-6): ___

Appendix E: Correlational Matric for All Numeric Variables in Logistic Regression Analysis for the Entire Dataset

	2	3	4	5	6	7	8	9
1. mean length of P-bursts	0.302	0.02	-0.4	-0.623	0.028	-0.059	0.449	0.253
2. mean length of R-bursts		0.037	-0.297	-0.102	-0.111	-0.479	0.164	0.163
3. mean length of pauses			-0.086	-0.122	-0.126	-0.159	0.239	0.11
4. mean length of within-word pauses				0.296	0.095	0.207	-0.12	-0.151
5. mean length of between-word pauses					0.061	0.085	-0.22	-0.102
6. mean length of deletions						0.854	0.223	0.157
7. proportion of deletions							0.151	0.065
8. typing speed								0.518
9. vocabulary knowledge								