Making sense of L2 written argumentation with keystroke logging

Yu Tian[1], Minkyung Kim[2] & Scott Crossley[3]

1 Georgia State University | USA

2 Korea National University of Education | South Korea

3 Vanderbilt University | USA

Abstract: This study examines associations between writing behaviors manifested by keystroke analytics and the formulation of argument elements in L2 undergraduate writers' writing processes. Ninety-nine persuasive essays written by L2 undergraduate writers were human annotated for Toulmin argument elements. The corresponding keystroke logs were segmented and analyzed to characterize the dynamics of writing processes for different categories of the elements. A multinomial mixed-effects logistic regression model was built to predict argument categories using the keystroke analytics. The study reported that L2 undergraduate writers' text production for final claims and primary claims featured P-bursts (execution processes delimited by pauses exceeding 2 seconds) of longer spans but lower production fluency compared to that for data. In addition, fewer revisions were observed when L2 writers were constructing final claims than when they were formulating data. These findings shed light on the varying cognitive loads and activities L2 undergraduate writers may experience when building different argument elements in written argumentation.

Keywords: L2 written argumentation, Keystroke logging, Cognitive activities

journal of
WRITING RESEARCH

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

Contact: Yu Tian, Georgia State University/ 25 Park Place, Suite 1500, Atlanta, GA 30303 | USA – ytian9@gsu.edu.

1. Introduction

Argumentation can be viewed as a logical appeal that involves stating claims and offering support to justify or refute beliefs in order to influence others (Jonassen & Kim, 2010; Newell et al., 2011). In academic settings, the ability to construct robust arguments is essential not only when writing generic, context-neutral academic essays for writing courses, but also when writing within different disciplines (Hirvela, 2017; Varghese & Abraham, 1998). Despite its importance, written argumentation has long been recognized as a challenging task for young adult learners (e.g., university students) due to its cognitively demanding nature (Wingate, 2012). Compared to their native English-speaking peers, second language (L2) learners may face an even bigger challenge in argumentative writing because of cross-linguistic differences in argumentation and linguistic proficiency. Indeed, research has documented that L2 learners tend to wrestle with potential interference from different cultural norms and writing practices when constructing arguments in English (Hirvela, 2017; Kubota, 2010; Lee, 2014).

Previous research in written argumentation has largely focused on identifying and analyzing the argument structures (e.g., claim, evidence, counterarguments, and rebuttals) represented in students' argumentative essays (e.g., Chandrasegaran, 2008; Nussbaum & Kardash, 2005; Nussbaum & Schraw, 2007). In recent decades, there has been an increasing interest in writers' text production processes as opposed to merely their written products (e.g., Conijn et al, 2022; Lindgren & Sullivan, 2019; Michel et al., 2020; Révész et al., 2022; Spelman Miller, 2000; Van Waes, & Leijten, 2015; Wengelin, 2006). Information about the writing process helps us better understand writers' behavioral patterns and cognitive activities in text generation.

Argumentation likely manifests itself in the writing process via different writing behaviors, which may provide a window into writers' associated cognitive activities. To illustrate, the writer's use of argument models or schemas might facilitate the overall control of idea organization in composing an argumentative text, which in turn frees up cognitive resources in working memory, accelerates the process of converting ideas into words (Favart & Coirier, 2006; McCutchen, 1996), and likely results in prolonged production bursts with fewer pauses in writing. However, few if any studies thus far have taken a closer look at how writers construct argumentation in text production through such process features as pauses and revisions.

The paucity of process-based accounts of written argumentation might be partially due to the difficulties in observing and analyzing the dynamics of text production (Sullivan & Lindgren, 2006). As writing on a computer becomes a norm in modern society, keystroke logging has gained increasing currency as an observational tool in writing research. Compared to other observation methods such as think-aloud and video recording, keystroke logging provides possibilities to capture the temporal details of every keystroke, cursor, and mouse movement during writing unobtrusively and ecologically (Lindgren & Sullivan, 2019). The large amount of fine-grained data created by keystroke logging allows for in-depth analyses of various writing behaviors in composing.

In this study, we use keystroke logging analytics to characterize L2 undergraduate writers' argument construction processes. By identifying the disparate argument components generated by the L2 writers and aligning them each to their corresponding keystroke logs, we aim to shed light to the links between keystroke logging analytics and the construction of specific argument categories in L2 writing process. This knowledge will contribute to our understanding of the different cognitive activities underlying L2 adult writers' argument development and in turn provide implications for writing instruction and assessment.

1.1 Argumentation

To conduct a comprehensive investigation into the argument structures in writing, it is necessary to employ a model of argumentation to identify the generic argument elements and treat them as basic units of analysis. Toulmin's (1958, 2003) model of argumentation, grounded in a theory of human argumentation and adaptive to various domains of argumentative discourse, has generally been recognized as an effective tool in capturing the type of organizational structures associated with argumentative writing (e.g., Ferretti et al., 2000; Nippold & Ward-Lonergan, 2010; Nussbaum & Kardash, 2005, Stapleton & Wu, 2015). Toulmin's model revolves around three main categories: *claim*, *data*, and *warrant*. According to Toulmin, a *claim* is an assertion in response to a problem. The basis for making a *claim* is derived from *data* (i.e., facts or observations about the situation under discussion). The link between a *claim* and *data* is authorized by a warrant that features common sense rules, laws, scientific principles, or thoughtfully argued definitions (Hillocks, 2011). Apart from these three key categories, Toulmin also added *qualifier*, *rebuttal*, and *backing* to capture different aspects in human reasoning. A *qualifier* is a modal term (e.g., *probably*, *presumably*) used to indicate the strength of the relationship between a *claim* and *data* conferred by the *warrant*. A *rebuttal* denotes circumstances in which the general authority of the *warrant* will not hold. Finally, the *backing* is the knowledge structure from which justifications for the *warrants* are derived.

While Toulmin's model provides an insight into the ways an argument is structured and the nature of justification in supporting claims (Bell & Linn, 2000; Jimenez-Aleixandre et al., 2000), the application of Toulmin's model to complex arguments as those seen in argumentative essays by adult writers is not without problems (e.g., Kunnan, 2010; Sampson & Clark, 2008; Wingate, 2012). For instance, the challenge in distinguishing between warrants and backing often leads to reliability issues (Crammond, 1998; Sampson & Clark, 2008). Potential ambiguity has also been reported in identifying claims and assembling relevant argument elements in Toulmin's scheme when applied to integrated arguments (Kunnan, 2010). Indeed, Wingate (2012) argued that Toulmin's model is less helpful when it comes to analyzing the large-scale structure or arguments at the macro level of argumentative essays.

There have been a few attempts to better identify argument structures in adult argumentative writing using modified versions of Toulmin’s scheme. Nussbaum and Kardash (2005), for example, adapted Toulmin's model to analyze argument structures commonly seen in college students' essays. Their modified Toulmin scheme helps to identify claims of different levels and types: the *final claim* (an opinion or a conclusion on the main question), *primary claims* (one or more reasons to support the final claim), *counterclaims* (potentially opposing opinions to the final claim), and *rebuttals* (claims used to refute the counterclaims). In their scheme, they also included *supporting reasons or examples* which can be used to back up the stated claims. Similarly, both Qin and Karabacak (2010) and Stapleton and Wu (2015) adopted a scheme based on Toulmin's model which comprised six elements: *claim*, *data*, *counterargument claim*, *counterargument data*, *rebuttal claim*, *rebuttal data*. In general, these modified versions, although varied in their specific focuses in argumentation analysis, have facilitated studies of argumentative discourses in adult writing and have provided valuable theoretical and methodological information.

1.2 Keystroke Logging

The application of keystroke logging in writing research as an observational tool for the writing process has been gaining momentum as computer-based writing has become increasingly prevalent. In general, current keystroke logging programs widely used in writing studies can record different keystroke operations such as insert, delete, cut, paste, and replace as well as mouse movements. Time stamps for these keyboard and mouse operations are reported to indicate when the events occur and how long they last. In some advanced keystroke logging programs such as Inputlog (Leijten & Van Waes, 2013), cursor position information can also be logged to allow for analyses of operations at different locations.

The detailed logs of keystroke activities make it possible to generate a myriad of measures related to the writing process including pauses, revisions, and bursts. For instance, the automatic calculation of the gap time between two consecutive key presses expressed in million seconds, or the so-called inter-keystroke intervals (IKI) (Chukharev-Hudilainen et al., 2019), provides an avenue towards multi-dimensional analyses of pause behaviors. In writing studies, it has been a common practice to define pauses as IKI above certain thresholds (e.g., 2 seconds). Pause behaviors can then be measured through indices of frequency and duration, and their locations in the text (e.g., within words, between words, between sentences) (e.g., Dich & Pedersen, 2013; Medimorec & Risko, 2016; Van Waes & Schellens, 2003; Van Waes et al., 2014). Keystroke logging can also keep tracks of online changes made throughout the text production process, such as the texts deleted, inserted and replaced, the time stamps and locations of the revision events, number of revisions, number of characters before and after revision, which supports analyses of different types of on-line revisions (see e.g., Conijn et al., 2022; Lindgren & Sullivan, 2006; Stevenson et al., 2006; Thorson, 2000; Van Waes & Leijten, 2015). Keystroke logs allow researchers to pinpoint and examine the periods in text production in which stretches of texts are continuously produced with no pauses and/or revisions. The fluent production of written language in temporal segments as such is referred to as "burst" (Kaufer et al., 1986). Two types of bursts have been distinguished in writing research: P-bursts that refer to the written segments terminated by pauses of over two or more seconds, and R-bursts that describe the segments terminated by an evaluation, revision, or other grammatical discontinuity (Chenoweth & Hayes, 2001).

1.3 Keystroke Logging and Cognitive Writing Processes

Skilled writing has been perceived as a conscious, demanding, and self-directed process, featuring a constellation of cognitive activities involving problem-solving and decision-making to satisfy the writer's goals (Kellogg, 1994). Some basic mental operations writers employ while composing, as described in a handful of influential writing process models (e.g., Flower & Hayes, 1981; Hayes, 1996), include *planning* to set goals and to generate and organize ideas, *translating* (or *formulating*) the ideas into linguistic strings, and *revising* to improve the text. The coordination of these processes is assumed to be governed by a *monitor*, and accordingly different configurations of the monitor are associated with different writing strategies in composing. There is also a control level which features task schemas, and a resource level which includes such important cognitive resources as long-term memory, working memory and attention that writers capitalize on in text production (Hayes, 2012). The constraints imposed by the writer's limited processing capacities and cognitive resources exert influence on their composing experience, compelling them to strategically allocate available resources to navigate the writing process (MacArthur & Graham, 2016). The interplay among these writing schemas, cognitive resources, and cognitive activities in composing likely manifest itself in the writer's various writing behaviors which can be captured using keystroke logging measures. Indeed, the majority of writing research using keystroke logging has taken a cognitive approach that maps keystroke units of analysis to specific components of the writing process with an aim to make inferences about the cognitive demands and processes in text production (e.g., Alves et al., 2008; Barkaoui, 2019; Chan, 2017; Chukharev-Hudilainen, 2014; Olive et al., 2009; Spelman Miller, 2000; Van Waes et al., 2014).

For instance, pause behaviors, as represented with IKIs in keystroke logging, are considered as important indicators of the writer's covert cognitive activities that are otherwise impossible to observe via the written products. Pause lengths are assumed to reflect cognitive activities of different levels. Although there is still debate on using certain pause thresholds to distinguish cognitive pauses (Galbraith & Baaijen, 2019), a pause above two seconds has been used as a rule of thumb in writing research to detect higher order cognitive processing such as planning for new ideas or revising (Chukharev-Hudilainen, 2014; Wengelin, 2006), whereas a pause between 30 milliseconds and two seconds may reflect transcription processes related to typographic skills and spelling (Limpo & Alves, 2017). In addition, pauses occurring at the boundaries of different linguistic units (e.g., characters, words, or sentences) are also assumed to reflect planning or decision-making processes at different textual domains (e.g., morphological, grammatical, or discourse levels). It is generally assumed that pause boundaries preceding higher linguistic units often indicate more demanding cognitive activities associated with planning and decision-making for production of larger chunks of text (Spelman Miller, 2000). Revising, as indicated by deleting and inserting activities in keystroke logs, has also been acknowledged as an integral cognitive component in the writing process models. It is assumed to be associated with discrepancies between the writer's intentions and the text generated so far (Leijten et al., 2010; Lindgren et al., 2008) and has also been documented by previous studies as an effortful process in text production (Kellogg, 1994; Stevenson et al., 2006). Information about what has been deleted or inserted in keystroke logging data can help distinguish revisions carried out at different levels, such as content revisions, surface language revisions, and typographic revisions. The length of bursts in text production is primarily perceived as an indicator of the writer's cognitive capacity in execution processes (Breuer, 2019). According to Galbraith and Baaijen (2019), P-bursts with a pause threshold of two seconds relate to the capacity of the translator in text production. On the contrary, R-bursts are related to episodes where evaluation has led to the termination of a burst before completion, and thus do not reflect the capacity of the translator.

Keystroke logging thus constitutes an important tool to capture different cognitive costs involved in written argumentation. Constructing and evaluating arguments in writing is a highly demanding problem-solving process that requires writers to tap into their mental resources (e.g., knowledge about the genre and topic) as well as use a set of self-regulatory strategies to relieve their cognitive load in composing (Ferretti & Fan, 2016). Given the complexity of argumentation in writing, the construction of different argument elements likely involves distinct mental activities and strategies, leading to varied degrees of mental effort. For example, Nussbaum (2008) suggested that different reasoning skills are involved in the construction of disparate arguments (e.g., claims, supporting reasons, counterarguments) in argumentative writing. Nussbaum and Schraw (2007) argued that writers often undergo high cognitive loads when considering counterarguments and integrating their own arguments and the counterarguments into an overall final position. Shehab and Nussbaum (2015) documented that college student writers reported different levels of mental efforts when employing different strategies to integrate arguments and counterarguments. However, to date, there has been a lack of comprehensive accounts of L2 writers' behaviors, writing strategies, and cognitive efforts in constructing specific argument elements based on inferences drawn from their keystroke activities in text production.

1.4 Current Study

This study examines links between writing behaviors manifested by keystroke analytics and the formulation of argument elements in L2 undergraduate writers' writing processes. The data we present contains a set of persuasive essays written by L2 undergraduate students with rich keystroke logging information by essay. The keystroke logging measures we use to characterize the L2 writers' writing behaviors include bursts, pauses, and revisions. The argument categories we use to describe the argument structures of these L2 essays are based on Toulmin's (1958, 2003) model of argumentation and comprise final claim, primary claim, counterclaim, rebuttal, and data. The research question is as follows:

> To what extent can keystroke logging measures predict argument categories in L2 argumentative writing?

2. Method

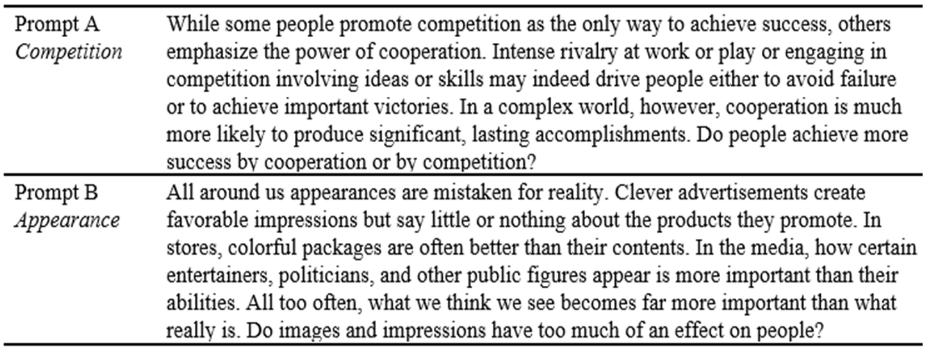
2.1 Participants

The data set used in this study was taken from Kim et al. (2021). In this study, 99 L2 undergraduate students at a research-oriented U.S. university each wrote one persuasive essay for which keystroke logging information was collected. Participant ages ranged from 18 to 31 years, with a mean age of 20.384 years (*SD* = 2.629). Among the participants, 66 were females. The participants were from a variety of linguistic backgrounds among which Chinese (*n* = 15), French (*n* = 15), and Spanish (*n* = 13) were the most common. They were also from a wide spectrum of academic disciplines, among which the three most popular were Science, Math & Technology (*n* = 34), Business (*n* = 26), and Social Science (*n* =23). In terms of the years of their study, there were 36 freshmen, 25 sophomores, 26 juniors, and 12 seniors. On average, these L2 undergraduate students had studied English for 12.869 years (*SD* = 4.082) by the time the data were collected.

2.2 Design and Procedures

Participants were invited individually to a soundproof language laboratory equipped with desktop computers. They were then asked to write one independent persuasive essay in 25 minutes on one of the two SAT-based prompts (see Table 1) in English. Two prompts were used to control for prompt effects. The prompts for the writing task were randomized among the participants such that 49 of them wrote on Prompt A (*Competition)* while 50 wrote on Prompt B (*Appearance*). Participants' keystroke activities during their writing were logged and time stamped via Inputlog 7 (Leijten & Van Waes, 2015).

Table 1. Two SAT-based Prompts



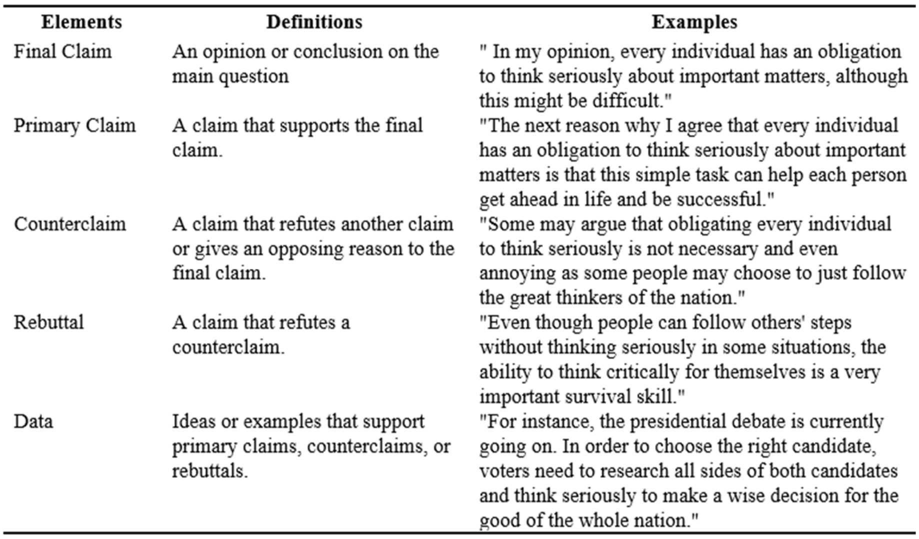
Prompt A *Competition*	While some people promote competition as the only way to achieve success, others emphasize the power of cooperation. Intense rivalry at work or play or engaging in competition involving ideas or skills may indeed drive people either to avoid failure or to achieve important victories. 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?
Prompt B *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?

2.3 Essay Annotation

2.3.1. Argumentation annotation rubric

To annotate the 99 persuasive essays, we adopted the argumentative rubric for classifying discourse elements used in Crossley et al. (2022). The rubric comprised five categories as the building blocks of the argumentation framework: *final claim*, *primary claim*, *counterclaim*, *rebuttal*, and *data*. This rubric was originally based on Toulmin's (1958, 2003) model but was modified based on work specific to argumentative writing as found in Liu and Stapleton (2014) and Nussbaum and Kardash (2005). Table 2 presents the definitions and examples for the argument categories.

Table 2. Definitions and Examples of Argument Elements (Crossley et al., 2022)



Elements	Definitions	Examples
Final Claim	An opinion or conclusion on the main question	" In my opinion, every individual has an obligation to think seriously about important matters, although this might be difficult."
Primary Claim	A claim that supports the final claim.	"The next reason why I agree that every individual has an obligation to think seriously about important matters is that this simple task can help each person get ahead in life and be successful."
Counterclaim	A claim that refutes another claim or gives an opposing reason to the final claim.	"Some may argue that obligating every individual to think seriously is not necessary and even annoying as some people may choose to just follow the great thinkers of the nation."
Rebuttal	A claim that refutes a counterclaim.	"Even though people can follow others' steps without thinking seriously in some situations, the ability to think critically for themselves is a very important survival skill."
Data	Ideas or examples that support primary claims, counterclaims, or rebuttals.	"For instance, the presidential debate is currently going on. In order to choose the right candidate, voters need to research all sides of both candidates and think seriously to make a wise decision for the good of the whole nation."

2.3.2 Annotation procedure

Two annotators were hired to annotate the L2 writers' persuasive essays using TagTog (https://www.tagtog.net), a web-based text annotation platform. Both annotators were PhD students who had experience teaching and coding argumentative writing. Before annotation, the two annotators were given two-hour's worth of instruction on the annotation rubric and the use of TagTog.

The annotators were then trained in a set of norming sessions wherein they independently annotated several batches of persuasive essays from another corpus (36 in total) on TagTog. Their annotation results were then compared and analyzed for differences by an expert annotator, who was a PhD student with over 10 years of experience teaching and researching writing structures and quality. The expert annotator helped to resolve disagreements and norm the annotators. After the norming sessions, the annotators independently coded the essays in opposite order to avoid recency effects.

For the annotations that resulted from the independent coding, we determined agreement using Jaccard Index to calculate overlap between the two sets of texts coded by the two annotators as one element (Tanimoto, 1958). The equation for the Jaccard Index is:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

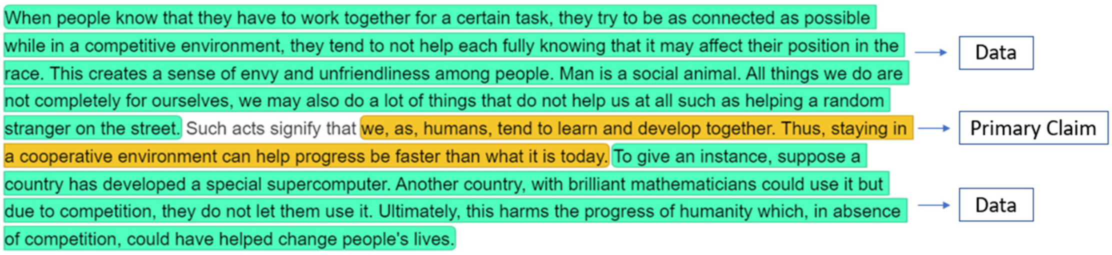
This index measures the size of the intersection between set A and set B (in this study, the number of characters tagged with the same category label in the two sets), divided by the size of the union of the two sets (the total number of characters tagged with the label in either set A or B). Agreement occurred when the Jaccard Index was more than .50 (i.e., there was an overlap of at least 50% in the two text sets tagged by the two annotators with the same category label). The average Jaccard Index between the two annotators for all their annotations was 0.617.

We then measured inter-annotator agreement (IAA) for the entire data set using Cohen's Kappa which represents the proportion of agreement in the annotation results after chance agreement is removed from consideration (Cohen, 1960). Cohen's Kappa is a recommended method to deal with annotations of linguistic data as used in this study with mutually exclusive categories without any ordering calculations (Bayerl et al., 2003). The Cohen's Kappa obtained was (k) = 0.651, $p <$ .001, suggesting substantial overall agreement between the two annotators (Landis & Koch, 1977).

To measure how well the two annotators agreed with each other when coding within each argument category, we employed the agreement formula found in Ferretti et al. (2009):

> Agreement Percentage = Number of Agreement / (Number of Agreement + Number of Disagreement)

Agreement percentages for the specific argument categories are presented in Table 3. Note that Table 3 does not include the categories of counterclaim and rebuttal. These categories were removed from the analyses because of their rarity in the data. For example, of the 410 annotations in the total dataset, only five were labeled as counterclaims and only four were labeled as rebuttals. Also of note is the relatively low agreement percentage between the two annotators when coding for primary claims. An examination of the annotators' coding results revealed that in annotation disagreements where a selected text area was coded as a primary claim by either of the two annotators, it is often the case that the other annotator coded a significant proportion (i.e., more than 50% characters) of the same text area as data. Specifically, among the 74 text sets tagged by annotator A as primary claims, a significant proportion of 30 sets were coded as data by annotator B. Among the 101 text sets labeled by annotator B as primary claims, 52 were identified as data by annotator A. The discrepancy between the two annotators when coding for primary claims is likely due to the variability of the argument sub-structures in the L2 writers' essays where primary claims were often implicitly introduced or closely interwoven with supporting reasons, facts, personal examples, etc., making it hard to distinguish them from data. This structural feature potentially led to disparate interpretations of the argument structures by different annotators. The following excerpt from one of the persuasive essays with adjudicated annotation results illustrated how a primary claim can be enveloped in the supporting details instead of being clearly presented at the beginning of the paragraph as a topic sentence:



When people know that they have to work together for a certain task, they try to be as connected as possible while in a competitive environment, they tend to not help each fully knowing that it may affect their position in the race. This creates a sense of envy and unfriendliness among people. Man is a social animal. All things we do are not completely for ourselves, we may also do a lot of things that do not help us at all such as helping a random stranger on the street. Such acts signify that we, as, humans, tend to learn and develop together. Thus, staying in a cooperative environment can help progress be faster than what it is today. To give an instance, suppose a country has developed a special supercomputer. Another country, with brilliant mathematicians could use it but due to competition, they do not let them use it. Ultimately, this harms the progress of humanity which, in absence of competition, could have helped change people's lives.

All the annotations by the two annotators were adjudicated by the third, expert annotator. Wherever there was a disagreement between the two annotators, the expert annotator adjudicated by making the final decision as to which annotation was more acceptable based on the rubric. In the present study, we used the adjudicated annotations in all analyses.

Table 3. Agreement Percentage for Each Argument Category

Argument Category	Agreement
Final Claim	0.82
Primary Claim	0.54
Data	0.91

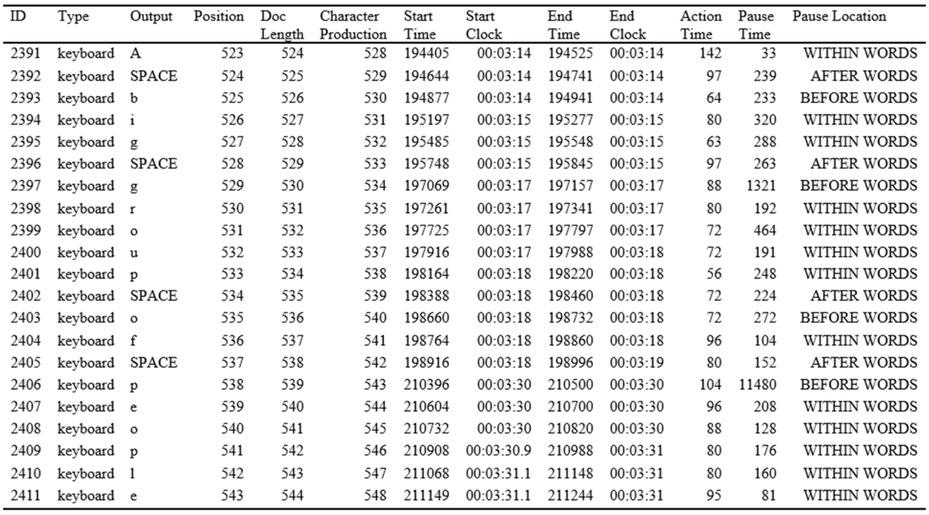
2.4 Keystroke Logging Measures for Each Argument Category

We obtained the keystroke logging information for each argument category produced in the L2 writers' writing process by employing the following procedure: First, we processed the keystroke logging files (records of the L2 writers' whole writing process) using Inputlog 7's built-in general analysis and saved the keystroke information in tabular dataframes where all keystrokes and their respective time stamps were arranged in a linear order (see table 4 for an example dataframe). Second, we analyzed the text evolvement in the L2 writers' writing process using the keystroke and position information in the spreadsheets and the restored text in larger units of sentence (or word units in some insertion activities). Third, we determined the timing information for the restored text segments by identifying their start and end time in text production. Fourth, we juxtaposed and compared the time stamped text segments with the argument annotation results to tag the text segments with their corresponding argument element labels (i.e., we obtained the timing information for each argument element produced by the L2 writers). Fifth, we split the keystroke logging files for each argument element according to their start and end time information in text production. Note that due to the nonlinearity of the text production process, a writer may not produce an array of linguistic strings consecutively for a specific argument element. Some L2 writers in this study, for example, were often seen diverting from their effort of building an argument element (e.g., data) at the point of inscription to revise or complete another element (e.g., primary claim) produced earlier in the text. This nonlinear nature of the writing process led to more segmented log files than the argument elements identified in the written products for some L2 writers.

Prior to analyzing the keystroke logging files, we merged the segmented files of the same argument category for each participant using Inputlog's merging function in its preprocess module. We did this for largely two reasons: 1) some segmented keystroke logging files only recorded a short episode of revision behavior (e.g., deletion or insertion) made to the previous text. In many cases, these log files lacked information on certain keystroke measures (e.g., there may be no information on pause between words). Hence, merging these log files with those of the same argument category helped reduce the number of missing values in the dataset and facilitate statistical analyses. 2) the focus of our study was to investigate whether argument categories can be predicted by keystroke analytics. Merging the discrete log files of the same argument category for each participant helped concentrate on the inquiry and minimize the random noise in the data resulted from writers producing more than one element of the same category.

We then analyzed the keystroke logging files (mapped to different argument categories) to generate a list of keystroke indices with reference to the L2 writers' P-bursts, pause behaviors, and revision activities. In this study, we operationalized P-bursts as continuous text production episodes terminated at pauses of two or more seconds given that this standard has been most commonly used in writing research to identify P-bursts (see e.g., Kaufer et al, 1986; Van Waes & Leijten, 2015). However, we set the threshold for pauses in the present study at 200 milliseconds because this threshold has the strength of capturing the bulk of language- and planning-related differences in pausing while filtering most inter-key intervals that result solely from the motor constraint of typing (Medimorec & Risko, 2017). We also analyzed pauses at different locations (i.e., within words, between words, between sentences) because they are associated with various patterns of pause behaviors and might provide insights into different underlying cognitive processes in writing (Chukharev-Hudilainen et al., 2019; Spelman Miller, 2000). As a result, a total of 28 keystroke logging indices were obtained to characterize the L2 writers' text production process.

Table 4. Examples of Output Generated by General Analysis in Inputlog 7



ID	Type	Output	Position	Doc Length	Character Production	Start Time	Start Clock	End Time	End Clock	Action Time	Pause Time	Pause Location
2391	keyboard	A	523	524	528	194405	00:03:14	194525	00:03:14	142	33	WITHIN WORDS
2392	keyboard	SPACE	524	525	529	194644	00:03:14	194741	00:03:14	97	239	AFTER WORDS
2393	keyboard	b	525	526	530	194877	00:03:14	194941	00:03:14	64	233	BEFORE WORDS
2394	keyboard	i	526	527	531	195197	00:03:15	195277	00:03:15	80	320	WITHIN WORDS
2395	keyboard	g	527	528	532	195485	00:03:15	195548	00:03:15	63	288	WITHIN WORDS
2396	keyboard	SPACE	528	529	533	195748	00:03:15	195845	00:03:15	97	263	AFTER WORDS
2397	keyboard	g	529	530	534	197069	00:03:17	197157	00:03:17	88	1321	BEFORE WORDS
2398	keyboard	r	530	531	535	197261	00:03:17	197341	00:03:17	80	192	WITHIN WORDS
2399	keyboard	o	531	532	536	197725	00:03:17	197797	00:03:17	72	464	WITHIN WORDS
2400	keyboard	u	532	533	537	197916	00:03:17	197988	00:03:18	72	191	WITHIN WORDS
2401	keyboard	p	533	534	538	198164	00:03:18	198220	00:03:18	56	248	WITHIN WORDS
2402	keyboard	SPACE	534	535	539	198388	00:03:18	198460	00:03:18	72	224	AFTER WORDS
2403	keyboard	o	535	536	540	198660	00:03:18	198732	00:03:18	72	272	BEFORE WORDS
2404	keyboard	f	536	537	541	198764	00:03:18	198860	00:03:18	96	104	WITHIN WORDS
2405	keyboard	SPACE	537	538	542	198916	00:03:18	198996	00:03:19	80	152	AFTER WORDS
2406	keyboard	p	538	539	543	210396	00:03:30	210500	00:03:30	104	11480	BEFORE WORDS
2407	keyboard	e	539	540	544	210604	00:03:30	210700	00:03:30	96	208	WITHIN WORDS
2408	keyboard	o	540	541	545	210732	00:03:30	210820	00:03:30	88	128	WITHIN WORDS
2409	keyboard	p	541	542	546	210908	00:03:30.9	210988	00:03:31	80	176	WITHIN WORDS
2410	keyboard	l	542	543	547	211068	00:03:31.1	211148	00:03:31	80	160	WITHIN WORDS
2411	keyboard	e	543	544	548	211149	00:03:31.1	211244	00:03:31	95	81	WITHIN WORDS

2.5 Statistical Analyses

We pruned the keystroke indices prior to statistical analyses. First, to facilitate comparison of keystroke activities between different argument categories, we excluded keystroke indices related with the absolute length of text or writing time and only retained indices that are based on means, proportions or ratios. Second, some keystroke indices in our dataset have a single value (e.g., 0 or 1) for the vast majority of the observations. These indices, also known as "near-zero variance variables", are commonly considered to have little predicative power in regression models (Kuhn & Johnson, 2013, p 44). To address this issue, we calculated 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) and filtered out the keystroke indices that had a low percentage of unique values (< 10%) and a high frequency ratio (> 20). Third, we conducted a series of correlation analyses among all the keystroke indices to assess multicollinearity between the indices. No pair of the keystroke indices was found highly collinear (absolute $r >$ .599). As a result, a total of six keystroke logging indices were retained for further analyses. These included *product process ratio* (i.e., the ratio of characters in the final product versus those produced in writing process), *mean length of P-bursts in seconds*, *mean length of P-bursts in characters*, *mean length of pauses*, *mean length of within-word pauses*, and *mean length of between-word pauses*. The definitions of these selected keystroke logging indices were presented in Table 5.

Table 5. Definitions of Keystroke Logging Indices Used in This Study

Keystroke Logging Indices	Definitions
product vs. process ratio	The number of characters in the product divided by the number of characters produced during the writing process.
mean length of P- burst in seconds	The mean duration of continuous text production delineated by an initial and end pause exceeding 2 seconds and is measured in seconds.
mean length of P- burst in characters	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 pauses	The mean length of latencies that exceed 200ms in text production and is measured in seconds.
mean length of within-word pauses	The mean length of latencies within words that exceed 200ms in text production and is measured in seconds.
mean length of between-word pauses	The mean length of latencies between words that exceed 200ms in text production and is measured in seconds.

To investigate whether keystroke logging measures were predictive of different argument elements produced in L2 undergraduate students' writing process, we performed a multinomial mixed-effects logistic regression. We entered the three-level argument category (*final claim*, *primary claim*, *data*) as the dependent variable. We chose data as a reference category for the response in order to perform a joint mixed-effects linear regression for the log-odds ratio of each category (in our case, final claim and primary claim) compared to the reference. Data is chosen as the reference category given its distinct rhetorical functions in argumentation compared to claims (Toulmin, 1958). The six keystroke logging measures were entered as fixed effects. We also included prompt (a categorical variable that has two values: *Appearance* and *Competition*) as a fixed effect to assess potential prompt effects in the production of argument elements (e.g., Knudson, 1992; Liu & Stapleton, 2014; Zhang, 1987). Participants were entered as the random effect which provided each participant with a unique intercept to quantify variation across them.

We performed the data analysis within a Bayesian framework using the R package *MCMCglmm* (Hadfield, 2010). One advantage of Bayesian statistics is that it allows for not only the test of the null hypothesis, as in traditional frequentist approach, but the estimation of the probability of specific parameter values given the data (Levshina, 2016). The Bayesian approach is also considered better suited for complex multilevel modeling and provides better interpretations of the results (McElreath, 2020). A salient feature of Bayesian statistics is that it requires the use of priors (i.e., the prior beliefs in the probability of specific parameters) in model fitting. The posterior probabilities of the parameters will then be estimated based on both the prior beliefs and the given data. Given the exploratory nature of this study, we opted for non-informative priors, which allows for an estimation of the posteriors based on merely the data. Another distinctive characteristic of the Bayesian approach is its compatibility with some advanced model fitting techniques such as Markov Chain Monte Carlo (MCMC) algorithm that directly draws samples from the posteriors to approximate the posterior distribution (Brooks, 1998). To achieve a sample size of 5000, we performed a total of 250,000 iterations for the Markov chains with a burn-in of 10,000 (i.e., the initial 10,000 MCMC iterations were discarded because initial samples might follow a very different distribution) and a thinning interval of 50 according to the formula provided by Levshina (2015).

MCMCglmm package provides two ways to check the relevance of the fixed effects. The first, which is typically Bayesian, computes the 95% credible interval (CI) of the parameter value and tests if this CI includes zero (the null value). Fixed effects whose CIs do not contain zero are considered significant. The other measure, which is more aligned with standard frequentist approach, derives a *p*-value of the mean posterior estimate for the parameter under the null hypothesis. In the model output, we included both measures.

3. Results

Descriptive statistics of the keystroke measures for each argument category are presented in Table 6. As shown, data elements were identified in all 99 essays. Final claims were produced in 93 essays and primary claims were produced in 52 essays. A pattern in the two P-burst measures is apparent in the descriptive data in which P-bursts in the formulation of primary claims and final claims were longer than in data but contained fewer characters. Additionally, the descriptive data seems to indicate that the product process ratio was higher in the production of final claims and primary claims than data. As for pause-related indices, slightly longer pauses were observed in text production for data than final claims and primary claims.

Table 6. Descriptive Statistics of Keystroke Analytics for Each Argument Category

	Final Claim (*n* = 93)		**Primary Claim** (*n* = 52)		**Data** (*n* = 99)	
Keystroke Logging Indices	*M*	SD	*M*	SD	*M*	SD
product vs. process ratio	0.865	0.083	0.844	0.098	0.822	0.07
mean length of P- burst in seconds	84.687	130.763	148.759	85.03	51.504	36.577
mean length of P- burst in characters	42.912	29.235	38.521	21.639	49.694	30.859
mean length of pauses	0.683	0.359	0.704	0.245	0.711	0.248
mean length of within-word pauses	0.373	0.098	0.367	0.078	0.387	0.102
mean length of between-word pauses	1.203	0.64	1.138	0.475	1.181	0.441

Our research question asks to what extent keystroke logging measures predict argument categories in L2 argumentative writing. Table 7 presents the results of the fixed effects in the regression analysis which include the posterior mean (similar to log-odds ratio in frequentist statistics) for each parameter value, the 95% CI, and the frequentist *p*-value. The argument category of final claim as compared to data was significantly predicted by *product process ratio* (p-MCMC < .000), *mean length of P-bursts in seconds* (p-MCMC < .000), and *mean length of P-bursts in characters* (p-MCMC < .004). The 95% CIs of these fixed effects indicated that the odds of final claims increased when there was a higher product process ratio, a longer P-burst in seconds, or a shorter P-burst in characters. The model reported no significant prompt effect on the production of final claims compared to data (p-MCMC = .988). There were also no significant differences in the L2 writers' pausing behaviors when producing final claims and data (p-MCMC = 0.142 for *mean length of pauses*, p-MCMC = 0.878 for *mean length of within-word pauses*, and p-MCMC = 0.167 for *mean length of between-word pauses*). The model also found that the argument category of primary claim as compared to data in writing was significantly predicted by *mean length of P-bursts in seconds* (p-MCMC < .000) and *mean length of P-bursts in characters* (p-MCMC < .000). Based on the CIs of these two fixed effects, these results indicated that the odds of primary claim increased when there was a longer P-burst in seconds or a shorter P-burst in characters. The writing prompt did not play a significant role in distinguishing between primary claims and data (p-MCMC = .264), nor did product process ratio (p-MCMC = .127). Additionally, no significant differences were found in the writers' pausing behaviors in composing primary claims and data (p-MCMC = 0.483 for *mean length of pauses*, p-MCMC = 0.237 for *mean length of within-word pauses*, and p-MCMC = 0.654 for *mean length of between-word pauses*).

Table 7. Fixed Effects of the Multinomial Mixed Effects Logistic Regression Model

	Posterior mean	**Lower boundary of 95% CI**	**Upper boundary of 95% CI**	**p-MCMC**
Final Claim vs. Data				
Intercept	-7.112	-11.852	-2.592	0.002**
Prompt: Competition	0.003	-0.73	0.722	0.988
Product process ratio	9.681	5.059	14.946	0.000***
Mean length of P-bursts in seconds	0.012	0.005	0.192	0.000***
Mean length of P-bursts in characters	-0.022	-0.039	-0.007	0.004**
Mean length of pauses	-1.18	-2.791	0.409	0.142
Mean length of within-word pauses	-0.306	-4.283	3.532	0.878
Mean length of between-word pauses	-0.601	-1.457	0.273	0.167
Primary Claim vs. Data				
Intercept	-2.088	-7.583	3.731	0.467
Prompt: Competition	0.528	-0.348	1.467	0.264
Product process ratio	4.383	-1.393	9.958	0.127
Mean length of P-bursts in seconds	0.018	0.011	0.025	0.000***
Mean length of P-bursts in characters	-0.038	-0.062	-0.016	0.000***
Mean length of pauses	-0.63	-2.466	1.022	0.483
Mean length of within-word pauses	-3.31	-8.816	2.045	0.237
Mean length of between-word pauses	-0.781	-1.912	0.262	0.154

* p<.05 ** p < .010 *** p< .001

We checked for issues of autocorrelation between successive draws in the Markov chains. No strong autocorrelation was detected, indicating that a next value was not influenced by the previous value in the chains. We also examined how well the Markov chains converged to the posterior distribution (i.e., reached stationarity) by visually checking the traces via the trace plots for each parameter. None of the plots shows a clear sign of bending in any specific direction, indicating good convergence (Lunn et al., 2013).

4. Discussion

The goal of this study was to investigate the associations between L2 undergraduate students' keystroke activities and the argument elements they constructed in the writing process. Unlike previous studies that only focused on the structural features of argumentation in the written products (e.g., Chandrasegaran, 2008; Nussbaum & Schraw, 2007), this study approached L2 written argumentation from a process-based perspective by first linking the writer's keystroke logs to individual argument categories and then building a multinomial mixed effects logistic regression model to predict the argument categories based on the keystroke analytics. In general, our study reported distinct keystroke behavioral patterns of L2 undergraduate writers when they composed argument elements of different categories.

Specifically, our study showed that when L2 undergraduate writers were constructing final claims compared to data in their text production, they tended to make fewer revisions (a higher product process ratio) and engage in longer durations of continuous text production between pauses (longer P-bursts in seconds) with fewer character produced (fewer characters per P-bursts). Taken together, these results generally indicate relatively lower cognitive costs for L2 writers while producing final claims in written argumentation. One contributing factor might be the initial planning performed before the composing process begins. Although we did not record this initial planning activity, it was plausible that writers were mentally formulating and rehearsing their stance on the issue described in the writing prompt while they were reading the prompt and getting prepared to write. This type of initial planning could be done rapidly, implicitly, and unconsciously (Torrance, 2016), thus more or less relieving the cognitive demand when they were actually writing the final claim. Another contributing factor might be related to the priming effect from the writing prompt. When formulating a final claim in an argumentative essay, the writer needs to advance his/her position or assertion in response to a contentious topic or problem as described in the writing prompt (e.g., "I agree that people achieve more success by cooperation."). In many cases, the writing prompt (as those used in the present study) provides alternative stances or possible solutions as well as essential linguistic cues to prompt the writer to formulate a final claim, which results in less demand on his/her cognitive capacity. In contrast, justifying claims with supporting data entails a set of complex reference, thinking, and reasoning skills (Schwarz & Asterhan, 2010). This is especially true in timed independent writing, as is the case in our study, where the writers may need to critically evaluate and analyze the situation while constantly retrieving from their long-term memory relevant knowledge and information to substantiate their claims, which constitutes high cognitive costs in writing.

The pattern of P-burst in the writing episodes for final claim as compared to those for data also merits explanation. In particular, the L2 undergraduate writers tended to write in longer durations of P-bursts in seconds when constructing final claims as compared to data. However, these prolonged P-bursts did not yield proportionally more text in terms of P-bursts in character. Instead, fewer characters were generated within the P-bursts when the L2 writers were constructing final claims. Whereas previous research generally documented the temporal and textual length of P-bursts as two congruent measures in characterizing P-burst (e.g., Revesz et al., 2017; Spelman Miller et al., 2008), our study suggested that this might not always be true, at least in L2 written argumentation. The contradicting results reported on these two measures of P-bursts (seconds versus characters) implied that the L2 undergraduate writers may have translated ideas during planning stages less fluently when formulating final claims compared to data. This might be attributed to the more linguistic constraints in constructing final claims given that the length of P-burst is commonly recognized as an indicator of individual linguistic capacities (Galbraith & Baaijen, 2019). Whereas L2 writers might have greater cognitive space when formulating supporting reasons, facts, and examples into the target language, they may presumably feel more rigidity in articulating the overarching final claim to meet the expectations of addressing the prompt head-on in clear and concise language.

The same P-burst pattern was reported for the production of primary claim as compared to that of data. The longer duration of P-bursts as displayed by the L2 writers in constructing primary claims might also be attributed to a planning effect. Writing behavior research has generally documented longer planning episodes and more content planning at the boundaries of larger linguistic units such as sentences or paragraphs (e.g., Linnemann, 2019; Spelman Miller, 2000; Schilperoord, 1996). Given that it is a common practice to formulate primary claims at the beginning of a paragraph to set the nature and scope of the ensuing argument, it may be postulated that L2 writers likely engaged in longer planning and developed fuller idea packages for primary claims than data, which resulted in a more prolonged execution period. However, the construction of primary claims in written argumentation is by no means easy, especially for non-expert writers. Primary claims are considered important nodes in argument substructures that contribute to structural complexity in extended argumentative text by connecting to the main claim and supporting evidence (Crammond, 1998). This feature might exert constraints on L2 writers' linguistic capacity in producing primary claims that are expected to not only echo the main points embodied in the final claim but also set the tone and scope of the supporting data. These constraints likely moderated the L2 writers' production rate when they were converting their ideas into words, resulting in fewer characters produced within P-bursts.

This study affords both theoretical and practical implications. Theoretically, this study demonstrated that writers' behaviors and strategies in argumentation can be observed and analyzed using keystroke logging techniques and the underlying cognitive activities involved can thereby be inferred. The results thus shed light on the interplay between writing behaviors and cognition in L2 written argumentation. This process-based approach to written argumentation may serve as a catalyst for follow-up studies that use keystroke logging to investigate a variety of other linguistic and rhetorical features related to argumentative writing. Practically, the revelation of this study that a set of keystroke logging measures significantly predicted the argument categories L2 writers constructed in the writing process point to a promising direction for current automated writing evaluation (AWE) development. A fundamental problem with the majority of AWE systems available today is that their feedback focuses only on the textual features of the written product and makes limited use of writing process data for feedback provision (Chukharev-Hudilainen, 2019). As can be extrapolated from our findings, integrating keystroke logging into AWE systems would help improve the systems' performance in discourse elements classification and in turn facilitate automated process-based feedback concerning writers' behaviors and strategies in written argumentation. Pedagogically, the different behavioral patterns in L2 written argumentation revealed in this study could be used by instructors to provide L2 students with more appropriate support programming. To illustrate, the P-burst patterns observed in L2 writers' formulation for final claims and primary claims may suggest the need for interventions in student planning strategies or language preparation for producing these elements. In this sense, keystroke logging data may help instructors to pinpoint L2 learners' difficulties in writing as a starting point for interventions.

5. Conclusion

In sum, the present study reported that L2 undergraduate writers displayed distinct keystroke behaviors when constructing different argument categories in writing argumentative essays. When constructing final claims compared to data, the L2 writers generally had fewer revisions and their writing process featured longer spans of P-bursts although their production fluency within these P-bursts was relatively lower. A similar writing behavior pattern was revealed when comparing their writing processes for primary claims and data, except that no significant difference was found in their revision behaviors. These findings shed light on the varying cognitive loads and activities L2 undergraduate writers may experience when building different argument elements in written argumentation.

One shortcoming of the dataset used in this study is the rarity of counterclaims and rebuttals found in L2 undergraduate writers' essays, which negated the opportunity for us to examine L2 writers' keystroke activities when they engaged in counter-argumentation. The paucity of counterclaims and rebuttals in our data is not surprising given that previous research has generally documented a lack of counterarguments in many L2 learners' argumentative writing. There are several possible contributing factors. Foremost, counter-argumentation likely entails increased cognitive load (Coirier et al., 1999) and a desire to maintain cognitive consistency (Simon & Holyoak, 2002), which might pose a challenge to most L2 writers. Moreover, many L2 writers might lack sufficient knowledge in counter-argumentation or they might be unfamiliar with the rhetorical traditions in Anglo-American academic essays where considering and rebutting an opposing side is often valued as a hallmark of good argument (Andriessen et al., 2003; Nussbaum & Schraw, 2007; Voss & Means, 1991). However, this problem can be mitigated by providing L2 writers with a clear goal instruction in writing prompts to elicit more counterarguments (Nussbaum & Kardash, 2005) or some short-term instructional interventions on counter-argumentation before testing their argumentation skills (Liu & Stapleton, 2014). Future research may use these strategies to increase the number of counterclaims and rebuttals in data collection.

There are also several other limitations that need to be acknowledged. Firstly, we did not take account of the L2 undergraduate writers' typing skills when analyzing their keystroke activities. However, individual variances in typing skills might have played a role in the keystroke outcomes given that these skills are considered an important factor affecting online text production when generating digital texts (Van Waes et al., 2019). Secondly, due to the scope of this study, we analyzed argumentative structures of the L2 essays by merely focusing on the presence of discrete Toulmin argument elements. Admittedly, the relationships between the elements as well as their relative positions or distances from one another in the text are also important structural features that were not analyzed (e.g., Crossley et al., 2022; Ferretti et al., 2009). Moreover, the sample size used in this study is relatively small, which might have constrained the statistical power of our model in predicting different argument categories. With a larger sample, more nuanced differences in keystroke activities of the L2 writers might have been detected and accordingly better classification results might have been yielded.

Overall, our study showed that a series of observed keystroke behaviors were associated with different argument categories L2 undergraduate writers constructed during the writing process. Inferences about the L2 writers' underlying cognitive activities while constructing these argument categories can be drawn based on the keystroke analytics. On a cautionary note, however, these inferences need to be taken conservatively. The alignment of keystroke measures with cognitive processes has never been unambiguous (Galbraith & Baaijen, 2019). While keystroke logging can be synchronous with the text production process, the logged keystrokes are considered to offer only indirect behavioral observation of the covert cognitive activities in a writing session. Hence, interpretations of the keystroke information in reference to the writer's cognitive processes are largely speculative and limited (Hoang, 2019). Although it is beyond the scope of our study to pinpoint the exact cognitive processes L2 writers undergo when constructing different argument elements, such a detailed, process-based cognitive account of L2 written argumentation is needed to help better understand how L2 writers coordinate different cognitive resources in written argumentation. A sensible practice in this concern would be to supplement keystroke logging data with eye-tracking, think-aloud protocols, or stimulated retrospective interviews to shed further light on writers' cognitive activities, as has been demonstrated in recent writing process research (e.g., Chukharev-Hudilainen et al., 2019; Michel et al., 2020; Leijten & Van Waes, 2013).

References

Andriessen, J., Baker, M., & Suthers, D. (2003). Argumentation, computer support, and the educational context of confronting cognitions. In J. Andriessen, M. Baker, & D. Suthers (Eds.), *Arguing to learn: Confronting cognitions in computer-supported collaborative learning environments* (pp. 1–25). Kluwer. https://doi.org/10.1007/978-94-017-0781-7

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

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. https://doi.org/10.1017/s027226311900010x

Bayerl, P. S., Lüngen, H., Gut, U., & Paul, K. I. (2003, October). Methodology for reliable schema development and evaluation of manual annotations. In *Proceedings of the Workshop on Knowledge Markup and Semantic Annotation at the Second International Conference on Knowledge Capture (K-CAP 2003).*

Bell, P., & Linn, M. C. (2000). Scientific arguments as learning artifacts: Designing for learning from the web with KIE. *International Journal of Science Education*, *22*(8), 797–818. https://doi.org/10.1080/095006900412284

Breuer, E. O. (2019). Fluency in L1 and FL writing: An analysis of planning, essay writing and final revision. In E. Lindgren & K. P. H. Sullivan (Eds.), *Observing Writing* (pp. 190-211). Brill. https://doi.org/10.1163/9789004392526_010

Brooks, S. (1998). Markov chain Monte Carlo method and its application. *Journal of the Royal Statistical Society: Series D (the Statistician), 47*(1), 69-100. https://doi.org/10.1111/1467-9884.00117

Chan, S. (2017). Using keystroke logging to understand writers' processes on a reading-into-writing test. *Language Testing in Asia, 7*(1), 1-27. https://doi.org/10.1186/s40468-017-0040-5

Chandrasegaran, A. (2008). NNS students' arguments in English: Observations in formal and informal contexts. *Journal of Second Language Writing, 17*(4), 237–254. https://doi.org/10.1016/j.jslw.2008.04.003

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

Chukharev-Hudilainen, E. (2014). Pauses in spontaneous written communication: A keystroke logging study. *Journal of Writing Research*, *6*(1), 61. https://doi.org/10.17239/jowr-2014.06.01.3

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. https://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. https://doi.org/10.1017/s027226311900007x

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

Coirier, P., Andriessen, J., & Chanquoy, L. (1999). From planning to translating: The specificity of argumentative writing. *Foundations of Argumentative Text Processing*, 1-28. https://doi.org/10.5117/9789053563403

Conijn, R., Speltz, E. D., Zaanen, M. V., Waes, L. V., & Chukharev-Hudilainen, E. (2022). A product-and process-oriented tagset for revisions in writing. *Written Communication*, *39*(1), 97-128. https://doi.org/10.1177/07410883211052104

Crammond, J. G. (1998). The uses and complexity of argument structures in expert and student persuasive writing. *Written Communication*, *15*(2), 230-268. https://doi.org/10.1177/0741088398015002004

Crossley, S., Tian, Y., & Wan, Q. (2022). Argumentation Features and Essay Quality: Exploring Relationships and Incidence Counts. *Journal of Writing Research*. https://doi.org/10.17239/jowr-2022.14.01.01

Dich, N., & Pedersen, B. (2013). Native Language Effects on Spelling in English as a Foreign Language: A Time-Course Analysis. *Canadian Journal of Applied Linguistics/Revue Canadienne de Linguistique Appliquee*, 16(1), 51-68.

Ellis, R. (2003). *Task-based language learning and teaching*. Oxford University Press.

Favart, M., & Coirier, P. (2006). Acquisition of the linearization process in text composition in third to ninth graders: effects of textual superstructure and macrostructural organization. *Journal of Psycholinguistic Research*, *35*(4), 305-328. https://doi.org/10.1007/s10936-006-9017-8

Ferretti, R. P., & Fan, Y. (2016). Argumentative writing. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (2nd ed., pp. 301–315). Guilford Press. https://doi.org/10.17239/jowr-2014.06.02.5

Ferretti, R. P., MacArthur, C. A., & Dowdy, N. S. (2000). The effects of an elaborated goal on the persuasive writing of students with learning disabilities and their normally achieving peers. *Journal of Educational Psychology*, *92*(4), 694. https://doi.org/10.1037/0022-0663.92.4.694

Ferretti, R. P., Lewis, W. E., & Andrews-Weckerly, S. (2009). Do goals affect the structure of students' argumentative writing strategies? *Journal of Educational Psychology*, *101*(3), 577. https://doi.org/10.1037/a0014702

Fisher, A. (2011). *Critical thinking: An introduction (2nd ed.)*. Cambridge University Press.

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

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. https://doi.org/10.1163/9789004392526_015

Godo, A. (2008). Cross-cultural aspects of academic writing: A study of Hungarian and North American college students LI argumentative essays. *International Journal of English Studies*, *8*(2), 65-111.

Hadfield, J. D. (2010). MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package. *Journal of Statistical Software*, *33*, 1-22. https://doi.org/10.18637/jss.v033.i02

Hayes, J. R. (1996). A new framework for understanding cognition and affect in writing. In M. Levy & S. Ransdell (Eds.), *The science of writing* (pp. 1-27). Erlbaum.

Hayes, J. R. (2012). Modeling and remodeling writing. *Written Communication*, *29*(3), 369–388. https://doi.org/10.1177/0741088312451260

Hillocks, G., Jr. (2011). *Teaching argument writing*, grades 6-12. Portsmouth, NH: Heinemann.

Hirvela, A. (2017). Argumentation & second language writing: Are we missing the boat? *Journal of Second Language Writing*, *36*, 69–74. https://doi.org/10.1016/j.jslw.2017.05.002

Hoang, H. (2019). Metaphorical language in second language learners' texts: Additional baggage of the writing journey? In E. Lindgren & K. P. H. Sullivan (Eds.), *Observing Writing* (pp. 236-257). Brill. https://doi.org/10.1163/9789004392526_012

Jimenez-Aleixandre, M., Rodriguez, M., & Duschl, R. A. (2000). "Doing the lesson" or "doing science": Argument in high school genetics. *Science Education*, *84*(6), 757–792. https://doi.org/10.1002/1098-237x(200011)84:6<757::aid-sce5>3.0.co;2-f

Jonassen, D. H., & Kim, B. (2010). Arguing to learn and learning to argue: Design justifications and guidelines. *Educational Technology Research and Development*, *58*(4), 439-457. https://doi.org/10.1007/s11423-009-9143-8

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

Kellogg, R. T. (1994). *The psychology of writing*. Oxford University Press.

Kim, M., Tian, Y., & Crossley, S. A. (2021). Exploring the relationships among cognitive and linguistic resources, writing processes, and written products in second language writing. *Journal of Second Language Writing*, *53*, 100824. https://doi.org/10.1016/j.jslw.2021.100824

Knudson, R. E. (1992). Analysis of argumentative writing at two grade levels. *The Journal of Educational Research*, 85(3), 169-179. https://doi.org/10.1080/00220671.1992.9944434

Kubota, R. (2010). Cross-cultural perspectives on writing. In N. H. Hornberger & S. Lee Mckay (Eds.), *Sociolinguistics and language education* (pp. 265-289). Multilingual Matters. https://doi.org/10.21832/9781847692849-012

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling.* Springer.

Kunnan, A. J. (2010). Test fairness and Toulmin's argument structure. *Language Testing*, *27*, 183–189. https://doi.org/10.1177/0265532209349468

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 159-174. https://doi.org/10.2307/2529310

Lee, S. H. (2014). Argument structure as an interactive resource by undergraduate students. *Linguistics & the Human Sciences*, *9*(3).

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

Leijten, M., Van Waes, L., & Ransdell, S. (2010). Correcting text production errors: Isolating the effects of writing mode from error span, input mode, and lexicality. *Written Communication*, 27(2), 189-227. https://doi.org/10.1177/0741088309359139

Levshina, N. (2015). *Bayesian logistic models with MCMCglmm: A brief tutorial.* Retrieved from http://www.natalialevshina.com/Documents/MCMCglmm_Tutorial.pdf

Limpo, T., & Alvès, R. A. (2017). Tailoring Multicomponent Writing Interventions: Effects of Coupling Self-Regulation and Transcription Training. *Journal of Learning Disabilities*, *51*(4), 381-398. https://doi.org/10.1177/0022219417708170

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). Information Science Reference, IGI Global. https://doi.org/10.4018/978-1-59904-895-6.ch011

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). Elsevier. https://doi.org/10.1163/9789004392526

Lindgren, E., & Sullivan, K. P. H. (Eds.). (2019). *Observing writing: Insights from keystroke logging and handwriting.* Brill. https://doi.org/10.1163/9789004392526

Linnemann, M. (2019). Application of audience during writing. In E. Lindgren & K. P. H. Sullivan (Eds.), *Observing Writing* (pp. 326-345). Brill.

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. https://doi.org/10.1016/j.system.2014.05.005

MacArthur C. A., & Graham, S. (2016). Writing research from a cognitive perspective. In C. A. MacArthur, S. Graham, S., & J. Fitzgerald, J. (Eds.), *Handbook of writing research* (pp. 24–40). NY: Guilford Press. https://doi.org/10.17239/jowr-2014.06.02.5

MacArthur, C. A., Graham, S., & Fitzgerald, J. (Eds.). (2008). *Handbook of writing research*. Guilford Press.

McCutchen, D. (1996). A capacity theory of writing: Working memory in composition. *Educational Psychology Review*, *8*(3), 299-325. https://doi.org/10.1007/bf01464076

McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan*. Chapman and Hall/CRC. https://doi.org/10.1201/9780429029608

Medimorec, S., & Risko, E. F. (2016). Effects of disfluency in writing. *British Journal of Psychology*, *107*(4), 625-650. https://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. https://doi.org/10.1177/0267658320915501

Newell, G. E., Beach, R., Smith, J., VanDerHeide, J., Kuhn, D., & Andriessen, J. (2011). Teaching and Learning Argumentative Reading and Writing: A Review of Research. *Reading Research Quarterly*, *46*(3), 273–304.

Nippold, M. A., & Ward-Lonergan, J. M. (2010). Argumentative writing in pre-adolescents: The role of verbal reasoning. *Child Language Teaching and Therapy*, *26*, 238–248. https://doi.org/10.1177/0265659009349979

Nussbaum, E.M. (2008). Using argumentation vee diagrams (AVDs) for promoting argument-counterargument integration in reflective writing. *Journal of Educational Psychology*, *100*(3), 549–565. https://doi.org/10.1037/0022-0663.100.3.549

Nussbaum, E. M., & Kardash, C. M. (2005). The Effects of Goal Instructions and Text on the Generation of Counterarguments During Writing. *Journal of Educational Psychology*, *97*(2), 157–169. https://doi.org/10.1037/0022-0663.97.2.157

Nussbaum, E. M., & Schraw, G. (2007). Promoting argument-counterargument integration in students' writing. *The Journal of Experimental Education*, 76(1), 59-92. https://doi.org/10.3200/jexe.76.1.59-92

Olive, T., Alves, R. A., & Castro, S. L. (2009). Cognitive processes in writing during pause and execution periods. *European Journal of Cognitive Psychology*, *21*(5), 758-785. https://doi.org/10.1080/09541440802079850

Qin, J., & Karabacak, E. (2010). The analysis of Toulmin elements in Chinese EFL university argumentative writing. *System*, *38*(3), 444-456. https://doi.org/10.1016/j.system.2010.06.012

Révész, A., Kourtali, N. E., & Mazgutova, D. (2017). Effects of task complexity on L2 writing behaviors and linguistic complexity. *Language Learning*, *67*(1), 208-241. https://doi.org/10.1111/lang.12205

Révész, A., Michel, M., Lu, X., Kourtali, N., Lee, M., & Borges, L. (2022). The relationship of proficiency to speed fluency, pausing, and eye-gaze behaviours in L2 writing. *Journal of Second Language Writing, 58*, 100927. https://doi.org/10.1016/j.jslw.2022.100927

Sampson, V., & Clark, D. B. (2008). Assessment of the ways students generate arguments in science education: Current perspectives and recommendations for future directions. *Science Education*, *92*(3), 447–472. https://doi.org/10.1002/sce.20276

Schilperoord, J. (1996). *It's about time. Temporal aspects of cognitive processes in text production*. Rodopi. https://doi.org/10.1163/9789004458598

Schwarz, B. B., & Asterhan, C. S. (2010). Argumentation and reasoning. *International handbook of psychology in education*, 137-176.

Shehab, H. M., & Nussbaum, E. M. (2015). Cognitive load of critical thinking strategies. *Learning and Instruction*, *35*, 51-61. https://doi.org/10.1016/j.learninstruc.2014.09.004

Simon, D., & Holyoak, K. J. (2002). Structural dynamics of cognition: From consistency theories to constraint satisfaction. *Personality and Social Psychology Review*, *6*(4), 283-294. https://doi.org/10.1207/s15327957pspr0604_03

Spelman Miller, K. S. (2000). Academic writers on-line: Investigating pausing in the production of text. *Language Teaching Research*, 4, 123-148. https://doi.org/10.1191/136216800675510135

Spelman Miller, K. S., Lindgren, E., & Sullivan, K. P. (2008). The psycholinguistic dimension in second language writing: Opportunities for research and pedagogy using computer keystroke logging. *TESOL Quarterly*, *42*(3), 433-454. https://doi.org/10.1002/j.1545-7249.2008.tb00140.x

Stapleton, P., & Wu, Y. A. (2015). Assessing the quality of arguments in students' persuasive writing: A case study analyzing the relationship between surface structure and substance. *Journal of English for Academic Purposes*, *17*, 12–23. https://doi.org/10.1016/j.jeap.2014.11.006

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. https://doi.org/10.1016/j.jslw.2006.06.002

Sullivan, K., & Lindgren, E. (2006). *Computer Key-Stroke Logging and Writing: Methods and Applications (Studies in Writing)*. Elsevier Science Inc. https://doi.org/10.1163/9780080460932

Tanimoto, T. T. (1958). *An elementary mathematical theory of classification and prediction*. Internal Report IBM Corp.

Thorson, H. (2000). Using the computer to compare foreign and native language writing processes: A statistical and case study approach. *Modern Language Journal*, 84(ii), 155–170. https://doi.org/10.1111/0026-7902.00059

Torrance, M. (2016) Understanding planning in text production. In C. A. MacArthur, S. Graham, S., & J. Fitzgerald, J. (Eds.), *Handbook of writing research* (pp. 72–87). Guilford Press.

Toulmin, S. (1958). *The uses of argument*. Cambridge: Cambridge University Press.

Toulmin, S. E. (2003). *The uses of argument*. Cambridge: Cambridge university press

Van Waes, L. V., & Leijten, M. (2015). Fluency in Writing: A Multidimensional Perspective on Writing Fluency Applied to L1 and L2. *Computers and Composition*, *38*, 79-95. https://doi.org/10.1016/j.compcom.2015.09.012

Van Waes, L., Leijten, M., 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. https://doi.org/10.5334/jors.234

Van Waes, L. & Schellens, P. J. (2003). Writing Profiles: The Effect of the Writing Mode on Pausing and Revision Patterns of Experienced Writers. *Journal of Pragmatics*, *35*(6), 829-853. https://doi.org/10.1016/s0378-2166(02)00121-2

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

Varghese, S. A., & Abraham, S. A. (1998). Undergraduates arguing a case. *Journal of Second Language Writing*, *7*, 287-306. https://doi.org/10.1016/s1060-3743(98)90018-2

Voss, J. F., & Means, M. L. (1991). Learning to reason via instruction in argumentation. *Learning and Instruction*, i(4), 337-350. https://doi.org/10.1016/0959-4752(91)90013-x

Wengelin, Å. (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). Elsevier. https://doi.org/10.1163/9780080460932_008

Wingate, U. (2012). 'Argument' helping students understand what essay writing is about. *Journal of English for Academic Purposes*, *11*, 145-154. https://doi.org/10.1016/j.jeap.2011.11.001

Zhang, S. (1987). Cognitive complexity and written production in English as a second language. *Language Learning*, *37*(4), 469-481. https://doi.org/10.1111/j.1467-1770.1987.tb00580.x