Argumentation in L1 and L2 Adult Writers: A Process-based Perspective Using Keystroke Logging

Argumentation in L1 and L2 Adult Writers: A Process-based Perspective Using Keystroke Logging

by

Yu Tian

Under the Direction of Scott Crossley, Ph.D. and Sara Cushing, Ph.D.

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

in the College of Arts and Sciences

Georgia State University

2023

ABSTRACT

Recent research using keystroke logging to investigate argumentative writing in assessment settings has documented associations between features of the writing process and writing quality. However, more research is needed to examine the direct links between the dynamics of the writing process and argument development. This dissertation investigated the interplay between behavioral patterns in the writing process as manifested in keystroke activities (i.e., pause, revision, burst, process variance), the formulation of argument elements (i.e., positions, claims, evidence, counterclaims, and rebuttals), and holistic scores of writing quality in L1 and L2 adult writers. Four hundred adult writers were recruited, including both L1 (n = 200) and L2 (n = 200) English speakers. Each wrote an independent persuasive essay on a randomly assigned SAT writing prompt. Their keystroke activities during text production were recorded. The essays were annotated for argument elements and human rated for holistic writing quality. For each writer, keystroke measures were developed for both the entire writing process and the formulation of each argument element. These measures were used in multinomial mixed effects logistic regression models to distinguish between argument element categories and in multiple linear regression models to predict writing quality. Results suggest that adult writers behaved differently in P-burst length, pause duration, and proportion of deletions when formulating different argument elements. Distinct keystroke patterns were also observed across L1 and L2 writer groups in formulating claims (compared to evidence) as well as counterclaims and rebuttals (compared to all other elements), indicating different cognitive loads the two groups underwent and different strategies they took in written argumentation. Last, for both L1 and L2 writers, higher writing scores were related to shorter pauses in general, shorter between-word pauses, lower proportion of deletions, higher proportion of insertions, and less process variance. This dissertation sheds light on the links between L1 and L2 adult writers' keystroke patterns in the writing process, the cognitive activities they engaged in, and the holistic quality of their written products. The dissertation affords important implications for writing research, writing instruction, and the development of intelligent writing support tools.

INDEX WORDS: Written argumentation, Keystroke logging, Writing quality, Cognitive models of writing, Adult writers, L2 writing

Argumentation in L1 and L2 Adult Writers: A Process-based Perspective Using Keystroke Logging

by

Yu Tian

Committee Chair: Scott Crossley

Sara Cushing

Committee: Luuk Van Waes

Diane Belcher

Electronic Version Approved:

Office of Graduate Services

College of Arts and Sciences

Georgia State University

August 2023

DEDICATION

To my wife, Linlin Duan, for her unwavering love, trust, and support.

ACKNOWLEDGEMENTS

This project would not have been possible without the guidance and support of many different people and organizations.

First and foremost, I am deeply indebted to my advisor and dissertation co-chair, Dr. Scott Crossley. Scott has offered guidance and encouragement at all stages of my academic career. At the beginning of my PhD journey, when I was plunged into the sea of information in language research, at a loss about how to navigate, it was Scott who timely introduced me to cutting-edge research methods and technologies that really expanded my repertoire of research skills. Since then, I have been venturing into different topics in writing research and experimenting with various novel methodologies (keystroke logging, natural language processing, and machine learning, etc.) under his sustained guidance. As a mentor, friend, and role model, Scott has taught me through his own example to keep an inquisitive mind and a conquering spirit in doing research. Another important lesson Scott taught me was: if you feel stressed out, go play badminton!

I would also like to express my deep appreciation to Dr. Sara Cushing, also my advisor and dissertation co-chair. Sara has given me tremendous encouragement and support during the tough time when I was juggling dissertation data collection, data analysis, and job hunting amid a series of other responsibilities. She was responsive and approachable whenever I needed her support. She made me at home in her office and always patiently and calmly listened to me report my new developments and provided valuable suggestions. I also really appreciate the time and effort she devoted to our Friday gatherings where I gained lots of encouragement and inspirations from discussing research ideas with colleagues and friends.

I want to express my gratitude to my dissertation members, Drs. Luuk Van Waes and Diane Belcher, for reading my dissertation drafts and offering constructive feedback. In particular, I would like to thank Luuk for his support when I was developing my own keystroke logging mechanism for data collection. He helped me better understand the technical standards used by Inputlog, provided guidance in testing the accuracy of my keystroke logging program, and made sure the output of my program aligned with that of Inputlog for data processing. I am also thankful to Diane for inspiring me to conduct research on argumentative writing and introducing me to various methods in genre analysis.

I also want to extend my special thanks to Dr. Ute Römer for serving as my custodian to supervise the use of grant money in data collection. Ute was really helpful in straightening out the confusing details in the budget. Her role was important because she saw to it that all the data were collected within budget.

I very much appreciate all the professors in the Department of Applied Linguistics and ESL at Georgia State University, especially those who taught the PhD courses I took. I would like to thank Dr. Youjin Kim for her profound knowledge and important contributions in second language acquisition research, which provided me with inspirations in analyzing argumentative texts by L2 writers in my dissertation. I also learned a lot from Dr. Stephanie Lindemann about a whole set of interesting topics and research methods in sociolinguistics.

Many thanks also go to all my friends and colleagues at Georgia State University. I am particularly grateful to my colleagues in both the department of Applied Linguistics & ESL and the department of English for choosing to work with me on my project in data collection as well as essay annotation and scoring: Andrew Howard Schneider, Anne Marie Giordano, Autumn Jones, Daria Potapenko, Elizabeth Topping, Maggie Miller, Penny Cook, Shabana Sayeed, Shuyuan Tu, and Tiffany Gray. I also want to thank my friends Andy, Joon, Jordan, Qian, Jan, Sanghee, Sally, Laura, and Yali for all of the venting sessions and good times.

In addition, I want to thank my wife, Linlin Duan, for her unwavering love and support in all ways (financially, emotionally, and particularly by cooking delicious food for every meal), and for her understanding and patience with me in the face of lost holidays and changed travel plans over the past few years.

Last but not least, I want to extend my sincere thanks to the organizations that funded my dissertation research: Schmidt Futures Foundation, Duolingo English Test, and The Adult Literacy Research Center (ALRC) at Georgia State University.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 INTRODUCTION

Argumentation is a logical appeal that generally involves stating claims and supporting the claims by evidence to justify or refute beliefs (Jonassen & Kim, 2010; Newell, Beach, Smith, & VanDerHeide, 2011). It is the means by which we rationally settle disputes, resolve issues of concern, and solve problems (Jonassen & Kim, 2009). Therefore, effectively arguing for a case has long been recognized as one of the key components of critical thinking and an essential skill in personal, professional, and academic contexts (Lee & Deakin, 2016; Pessoa, Mitchell, & Miller, 2017). Argumentation in writing is sophisticated and nuanced skill that entails intensive training and practice (Applebee, 1984; Liu & Stapleton, 2014). In educational settings, for instance, the ability to formulate strong arguments in academic language is crucial when writing generic, context-neutral academic essays for writing courses as well as when writing within different disciplines (Hirvela, 2017; Varghese & Abraham, 1998).

Despite its importance, written argumentation has long been recognized as a challenging task for many adult writers due to its cognitively demanding nature (Wingate, 2012). Compared to their native English-speaking peers, second language (L2) writers who grow up speaking a language other than English may face an even bigger challenge in argumentative writing because of a lack of target language mastery and cross-linguistic differences in argumentation. Foremost, the success of constructing effective written arguments strongly depends on the availability and accessibility of linguistic means necessary to framing clauses and sentences in grammatically correct and pragmatically adequate ways (Schoonen, Snellings, Stevenson, & van Gelderen, 2009). In this sense, L2 writers with inadequate target language proficiency may overburden their working memory with vocabulary searches and morphosyntactic consideration, which will limit their attentional resources allocated to higher-level cognitive activities (e.g., idea generation, evaluation) as those required in written argumentation (Kellogg, 1994; McCutchen, 1996). Therefore, writing in an L2 is generally much harder and more time consuming than in the native language (Chenoweth & Hayes, 2001). Furthermore, research into cross-linguistic differences in written argumentation has generally pointed to L2 writers' lack of familiarity with the rhetorical conventions and norms in English argumentative writing (Lee, 2014). To compound the situation, the argument schema (i.e., a structure that represents extended stretches of argumentative discourse; Anderson et al., 2001) many L2 writers have acquired from their L1 experience may run counter to what they are expected to do following English writing conventions. Consequently, many L2 writers have to wrestle with potential interference from different cultural norms and writing practices when constructing arguments in English (Hirvela, 2017; Kubota, 2010).

Previous research in written argumentation has largely taken a product-based approach (Crammond, 1998) and has focused on identifying the argument structures (e.g., claim, evidence, counterarguments, rebuttals) represented in writers' argumentative essays (e.g., Chandrasegaran, 2008; Nussbaum & Kardash, 2005; Nussbaum & Schraw, 2007). However, research into the process of written argumentation has been relatively rare despite the long tradition and an ever-growing interest in studying writers' text production processes as opposed to merely their written products (see e.g., Lindgren & Sullivan, 2019; Spelman Miller, 2000; Van Waes, & Leijten, 2015; Wengelin, 2006). Information about the writing process helps us understand writers' behaviors and cognitive activities while generating text. These activities can be important for assessing writer focus, the revision process, and content development.

It is arguable that argumentation may manifest itself in the writing process via different writing behaviors (e.g., planning, pausing, revising) which could potentially provide a window into the cognitive activities involved in argumentation. To illustrate, counterargumentation likely entails high cognitive demand (Nussbaum & Schraw, 2007) that competes for cognitive resources in working memory. This might leave traces in writers' text production processes characterized with pausing and revising behaviors. In addition, the use of argument models or schemas pertaining to the overall control of idea organization in composing an argumentative text might help free up cognitive resources in working memory and thus facilitate the process of converting ideas into words (Favart & Coirier, 2006; McCutchen, 1996). However, few if any studies thus far have taken a closer look at how writers construct their argumentation in text production through such process features as pauses and revisions.

In recent years, keystroke logging tools have increasingly been used to observe the writing process within the writing research community. 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 such as pauses and revisions in composing. Keystroke logs of argumentative writing processes thus serve as a highly promising source of data to measure and investigate how writers construct arguments from a process-based perspective.

This dissertation attempts to use keystroke logging techniques to investigate associations between the dynamics of the writing process, argument construction, and writing quality in L1 and L2 adult writers. There are three overarching purposes of this dissertation: a) investigate the links between keystroke features of writing process and the construction of different argument elements in L1 and L2 adult writers; 2) compare L1 and L2 adult writers in terms of their behavioral patterns in argument construction; and 3) examine the relationships between keystroke features and holistic writing quality in these two groups of writers.

This dissertation sheds light on the interplay between adult writers' behavioral patterns as manifested in their keystroke activities, argument construction, and the writing outcomes. In particular, examining how L1 and L2 writing behaviors vary in the formulation of different argument elements can help provide valuable inferences of the cognitive activities underlying written argument development. Furthermore, by comparing L1 and L2 writers in their behavioral patterns while generating different argument elements, this dissertation contributes to the currently under-researched scholarship on the differences between L1 and L2 written argumentation. Lastly, by investigating the relationships between process features and writing performances in argumentative writing, this dissertation helps deepen our knowledge about how different writing behaviors and strategies (e.g., planning, revising, time allocation) lead to different writing outcomes. The process-based approach to written argumentation used in this dissertation 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. The dissertation also has practical and pedagogical implications. For example, the findings of this dissertation may inform current automated writing evaluation (AWE) development because current AWE feedback focuses only on the textual features of the written product and not the process (Chukharev-Hudilainen, 2019). Integrating keystroke logging into AWE systems may allow for feedback at discourse level that is more accurate and pertinent as well as process-based evaluations or real-time feedback which could accommodate learners' needs during writing. Moreover, the dissertation has implications for argumentative writing instruction in general. By unraveling dynamics in L1 and L2 written argumentation processes, this dissertation may help instructors and administrators provide both L1 and L2 adult writers appropriate support programming and services based on a better understanding of their strengths and weaknesses related to written argumentation.

2 LITERATURE REVIEW

In this section, relevant previous research is presented according to five themes: (a) written argumentation; b) keystroke logging; c) keystroke logging and cognitive processes in written argumentation; d) keystroke features and argumentative writing quality.

2.1 Written Argumentation

2.1.1 Conceptualizing argumentation with Toulmin's model and its variants

A starting point in analyzing argument structures in argumentative text is perhaps to define the generic components that constitute an argument. In writing scholarship, different theoretical assumptions have been proposed to conceptualize and identify arguments (e.g., Andrews, 1995; Evensen, 2002; Hillocks, 2011; Newell et al., 2011; Reznitskaya & Anderson, 2002; Toulmin, Reike, & Janik, 1984; Van Eemeren, Grootendorst, & Henkemans, 1996; Walton, 1996). Among these frameworks, Toulmin's (1958, 2003) model of informal argument has been proved useful in capturing the nuances in the organizational structures of various argumentative texts (e.g., Connor, 1990; Ferretti, MacArthur, & Dowdy, 2000; Ferris, 1994; Knudson, 1992; McCann, 1989; Nippold & Ward-Lonergan, 2010; Nussbaum & Kardash, 2005, Qin & Liu, 2021; Stapleton & Wu, 2015; Yang, 2022).

Grounded in a theory of human argument, Toulmin's model of argument serves as a field-independent paradigm that presents a universal system of norms to describe the logical microstructure of arguments (Toulmin, 1958). Hence, Toulmin's model is deemed well-suited to the analysis of argumentative discourse of a variety of domains and purposes. There are three kernel categories in Toulmin's model, namely, claim, data, and warrant. The claim, the basic purpose of an argument, is usually an assertion or thesis to declare what exists or what values people uphold. The category of data provides grounds or foundation to support the claim, usually in the form of specific facts, evidence, or observations of the situation. The warrant, implicitly or explicitly stating the common-sense principles accepted as laws, scientific rules or studies, or thoughtfully couched definitions, serves as a bridge between the claim and its supporting data (Hillocks, 2011). Apart from these three core categories (i.e., claim, data, and warrant) that make up the outline of an argument, Toulmin also included three additional argument elements (i.e., qualifier, rebuttal and backing) to flesh out the model by capturing different aspects in the nature of human reasoning. Although not indispensable, these three elements are necessary in constructing a decent argument. Specifically, the qualifier is used in the wording of the claim in the form of linguistic signals (e.g., probably, likely) to modify the claim or convey the strength of the link between the claim and its supporting data conferred by the warrant. The rebuttal acknowledges circumstances where the claim might be invalid or the general authority of the warrant would have to be set aside. Last, the backing is the knowledge structure that provides justifications for the warrant and establishes its relevance. Without backing, the warrant may lack authority or currency in the argument. See Figure 2.1 for an illustration of how the six argument elements (each instantiated with an example provided by Toulmin) are presented in relation to each other in the argument paradigm.

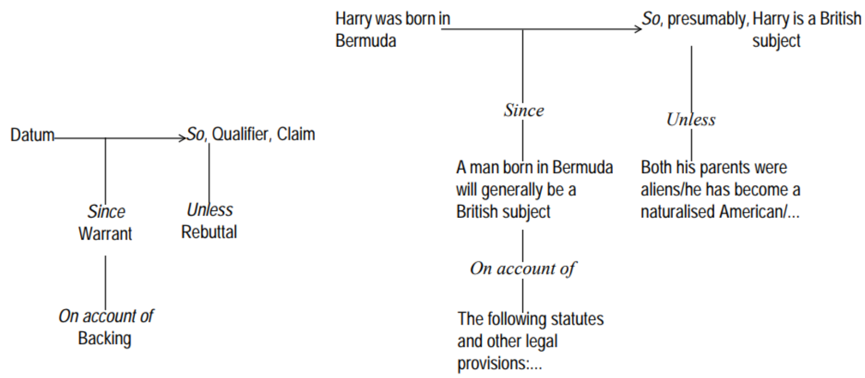

Figure 2.1 Toulmin's Model of Argument

Although Toulmin’s model has been effective in describing the ways a typical argument is structured and the nature of justification in supporting claims (Bell & Linn, 2000; Jimenez-Aleixandre, Rodriguez, & Duschl, 2000), it generally falls short in capturing the whole picture of human reasoning when applied to real, complex arguments as those seen in argumentative text by adult writers. First and foremost, reliability issues may arise when Toulmin's model is employed to identify different argument elements (e.g., claim, data, warrant, qualifier, rebuttal, backing). Sampson and Clark (2008) pointed out that one complication researchers often encounter when using Toulmin's model involves reliably identifying Toulmin elements in argumentative discourse. They provided a sample argument where some statements in the argument can serve as either a part of the claim, or a qualifier, or even a rebuttal. This issue may get more complicated when Toulmin's framework is applied to longer arguments as those in journal articles or course papers because some statements can be classified as sub-claims or warrants for a preexisting claim. According to Sampson and Clark, researchers' personal views and interpretations of arguments may influence how they code the arguments using Toulmin's analytic framework. They hence raised the alarm for potential personal biases when applying Toulmin's model to argument analysis in argumentative texts and the adverse effect they may have on interrater reliability. Likewise, Simon (2008) also noted the challenge of classifying

Toulmin elements in argumentation discourse by stressing that sometimes claims need to be deduced because of their implicit presentation in the argumentative text. The issue might be further confounded by the ambiguous nature of data, warrants, and backings in arguments. The same ambiguity issue was also raised by Kunnan (2010) who pointed out that classificatory confusion may arise when dealing with explicit and implicit claims coexisting in argumentative discourse. Kunnan also reported the complication in pinpointing claims and assembling related argument elements when applying Toulmin's framework to integrated arguments. He argued that the six argument elements in Toulmin's model "carve up the main argument into separate categories, thus failing to provide the synthesis needed to defend a claim" (p. 187). Moreover, although Toulmin's framework may be well-suited to interpreting the nature of human reasoning within a single argument, it is less effective to accommodate the large-scale structure or the macro-level of complicated arguments as is the case in extended argumentative discourse (Wingate, 2012).

Given the limitations of Toulmin's model in analyzing argument structures as those in argumentative texts, many researchers employed modified versions of Toulmin’s framework to attend to the multi-level arguments in argumentative essays in an attempt to yield more reliable classification of argumentative components (Nemeth & Kormos, 2001; Nussbaum & Kardash, 2005; Nussbaum & Schraw, 2007; Qin & Karabacak, 2010; Qin & Liu, 2021; Stapleton, 2001; Varghese & Abraham, 1998). For instance, Nussbaum and Kardash (2005) developed a modified version of Toulmin's model when analyzing argument structures in argumentative essays written by U.S. undergraduate students. This modified version captured the essential elements commonly seen in college students' essays including the final claim, the primary claim, the counterclaim, the rebuttal, and the supporting reasons or examples. Specifically, the final claim states an opinion or a conclusion on the main question which can be supported by one or more reasons functioning as the primary claim (s) in the text. The counterclaim raises potential opposing opinions to the final claim whereas the rebuttal is generally used to refute the counterclaim to further justify the final claim. Last, the supporting reasons and examples are evidence constructed to support the different claims stated above. A similar framework based on Toulmin's model was also used by both Stapleton and Wu (2015) and Qin and Liu (2021) which comprises six components: claim, data, counterargument claim, counterargument data, rebuttal claim, rebuttal data. In sum, these modified versions of Toulmin's scheme, although varied to some degree depending on the different focuses in their analyses, generally facilitate more pertinent examination of argumentative discourses in different contexts and help derive valuable theoretical and methodological information in written argumentation research.

2.1.2 Comparing L1 and L2 written argumentation in English

Prior to presenting previous literature on L1 and L2 written argumentation in English, it is necessary to acknowledge all the complications involved in the prevailing bilateral view of L1 versus L2 in writing scholarship. Although the term "second language" or "L2" in the area of language research usually refers to any non-native language in a speaker's repertoire (Hammarberg, 2014), and accordingly an L2 writer is commonly defined as someone born and raised where the primary language used by parents or other adult caregivers in the family is not the L2 (Ferris & Hedgcock, 2014), the composition of the L2 writer group following these broad definitions is far from homogeneous. According to Ferris and Hedgcock (2014), the L2 English writer population can be comprised of four main sub-groups: 1) international students who study in an English-speaking country on student visas and whose home language is not English; 2)

EFL students who learn English as a foreign language (FL) in their home contexts; 3) resident immigrants who have relocated to the new L2 context where English is the dominant language; 4) generation 1.5 immigrants who are children of first-generation immigrants and who have received all or part of their formal education in the new country, particularly the U.S.A. (Harklau, Losey, & Siegal, 1999). Therefore, the L2 English writer group may vary widely in terms of their English language proficiency. Despite the dispute and confusion revolving around the L1/L2 dichotomy in writing research, this dissertation follows the broad view of L2 writers as described in Ferris and Hedgcock (2014) for an extensive literature search on written argumentation.

Previous research has generally documented written argumentation as intrinsically a challenging task, and this challenge manifests itself as much greater for L2 writers in comparison to their L1 counterparts. In particular, many L2 writers tend to show a relatively underdeveloped argumentation capacity. For example, Connor (1984) analyzed the structures of English argumentative essays (N = 6) written by two L2 college students (one is a native speaker of Japanese and the other Spanish) randomly selected from an advanced ESL class at a U.S. university. When comparing these L2 essays with those by native speakers, Connor concluded that L2 writers generally lacked justifying support following claims to build the argument. Likewise, Choi (1988) examined the text structure in English argumentative essays (11 essays on 3 argumentative topics) written by Korean speakers and L1 English speakers and reported that whereas all L1 English essays consistently included the elements of claim, justification, and conclusion, these elements were sometimes incomplete in the L2 learners' essays. Ferris (1994) compared the rhetorical strategies used in 60 persuasive essays written by L2 undergraduate students and their L1 English speaking counterparts. By analyzing the rhetorical variables presented in the essays using Toulmin's model, Ferris found that L2 writers were less likely to produce Toulmin argument components (i.e., claim, data, and warrant) and included significantly fewer counterarguments in their persuasive essays than the L1 writers.

Some more recent studies have also confirmed a general deficiency in argumentation with L2 writers of English. For instance, Cheng and Chen (2009) employed Toulmin’s model of argument to examine the use of English argumentation features in essays written by Taiwanese (*n* = 40) and US college freshmen writers (*n* = 39). The results showed that although both groups of writers supported their opinions with reasons to formulate an argument, significantly fewer arguments (argument size) and less diversity of argument structures (number and variety of elaboration structures) were found in Taiwanese writing. Specifically, U.S. writers outnumbered Taiwanese students in the use of data, warrants and countered-rebuttals. Compared to their U.S. peers, Taiwanese writers were less capable in elaborating their ideas by utilizing backing to formulate argument chains in support of other argument structures. They were also found less adept in applying warrants to justify the links between their reasons and claims. Qin and Karabacak (2010) analyzed text structures of argumentative essays written by L2 university students (*n* = 133) in China based on an adapted Toulmin's model comprising six elements (i.e., claim, data, counterargument claim, counterargument data, rebuttal claim, and rebuttal data). They reported that although argumentative essays written by these L2 university students included one claim and four pieces of data on average, a large proportion of these essays failed to address the opposing views and thus contained no counterargument- and rebuttal-related claims and data. Qin and Karabacak claimed that the lack of counterarguments and rebuttals in these L2 writers' essays led to the weakness in their argumentation. Stapleton and Wu (2015) examined English argumentative essays written by 125 high school students in Hong Kong using a modified Toulmin model that included claims, counterargument claims, rebuttals and their associated data. The study detected a large proportion of the essays without inclusion of counterarguments and rebuttals. In addition, unlike prior studies that mainly focused on the surface argument structures (i.e., the existence of Toulmin-like elements in the argument), Stapleton and Wu also assessed the quality of argument in six exemplary essays that used all the argument elements, and concluded that despite organizing their argumentation in a model fashion, these L2 writers often included weak, unacceptable or irrelevant reasons.

Similar results were also found when comparing argumentative essays written by the same nonnative English-speaking individuals in both their native languages and English. For instance, Yu and Atkinson (1988) investigated essays written by students (n = 124) from an English-medium secondary school in Hong Kong in both their L1 (Chinese) and L2 (English), and documented less effective linking of arguments in the English essays than the Chinese essays. In a similar vein, Kamel (1989) conducted a quasi-experimental study in which 44 Arabic adult learners of English composed argumentative essays in both Arabic and English. Using Toulmin's argument model as one of the frameworks for analysis, Kamel reported that the Arabic-speaking writers included fewer claims, data, and warrants in their English essays compared to those written in their native language. Similar findings were reported in Cheng and Chen (2009) where Taiwanese college writers' English essays were examined and contrasted against the essays written in their native tongue (Chinese). The results indicated that Taiwanese writers employed a lower number of arguments and reasons to strengthen claims while writing in English. They were also found less capable to diversify their arguments with various types of argument structures when writing in English than in Chinese.

Taken together, the evidence of L2 writers' inadequate argumentation skills as reported in previous L2 writing research substantiated the struggles these nonnative writers faced when dealing with the rhetorical aspects of English argumentative writing. The challenge of constructing strong written arguments for many L2 writers can be arguably attributed to the difficulty and complexity of producing argumentative discourse in itself (Bereiter & Scardamalia, 1982). Nevertheless, compared to their L1 English-speaking counterparts, L2 writers are also confronted with additional linguistic constraints in writing (Olson, Scarcella, & Matuchniak, 2013). Indeed, many L2 writers might lack sufficient knowledge concerning the linguistic features that typify and often signal rhetorical moves in argumentation. Another contributing factor might pertain to the different rhetorical or argumentation traditions L2 learners have acquired from their L1 literacy experiences. Cross-cultural studies on writing and argumentation thus far have provided evidence for the existence of different argument patterns across cultures, such as inductive vs. deductive presentation of arguments (e.g., Hirose, 2003; Kobayashi & Rinnert, 2008; Kubota, 1998; Uysal, 2008), the use of hedging in argumentation (e.g., Hazen, 1986; Hinkel, 1997; Hyland, 1996), and the type of evidence (e.g., Hazen, 1986; Kamimura & Oi, 1998; Okabe, 1983). Although argumentation practices are not necessarily contradictory across cultures, potential interference may still arise when L2 writers are expected to follow the Aristotelian rhetoric as the preferred mode of arguments in English writing norms while downplaying the importance of their L1 rhetorical conventions.

2.2 Keystroke Logging

2.2.1 Observing writing processes with keystroke logging

Recent decades have seen a surge of interest in the study of what the writer does (e.g., planning, pausing, revising) in the writing process as opposed to what the final product looks like (e.g., vocabulary, grammar, conventions) within writing research and pedagogy (see e.g., Lindgren & Sullivan, 2019; Spelman Miller, 2000; Van Waes, & Leijten, 2015; Wengelin, 2006). Information about text production processes provides avenues to help us better understand writers' behavioral actions and cognitive activities while generating the text. However, the observation and analyses of the dynamics of writing process have always been a challenging issue (Sullivan & Lindgren, 2006). Some long-established techniques and methods commonly used for observing writing processes such as think-aloud protocols (e.g., Hayes & Flower, 1980; van den Bergh & Rijlaarsdam, 2007) and retrospective reports (e.g., Lindgren & Sullivan, 2003; Schumacher, Klare, Cronin, & Moses, 1984) have proved effective in helping elicit information about writers’ cognitive processes. However, these methods are not without limitations in terms of reliability and validity. For instance, the think-aloud approach requires writers to spontaneously verbalize their thinking as they write. This concurrent self-report approach is considered to be intrusive and potentially interfere with or disrupt writers' cognitive processes in text production (Russo, Johnson, & Stephens, 1989). The non-intrusive retrospective self-report method provides rich information about thought processes during writing, but it relies on the writers’ memory to reconstruct their experiences (Wengelin, Frid, Johansson, & Johansson, 2019). More recently, screen-capturing has also been used to help unobtrusively observe and record the writing processes without requiring writers’ additional efforts in reporting their cognitive processes during or after their writing. However, this method is better suited for pedagogical feedback rather than granular analyses of writing process features to understand writers’ behavioral patterns (e.g., Brick, & Holmes; Hamel & Séror, 2016; Stannard, 2019).

In recent years, more attention has been drawn to keystroke logging as an observational tool in writing research. Compared to other observation methods, keystroke logging provides possibilities to capture the temporal details of every keystroke, cursor, and mouse movement in writing processes 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 such as pauses and revisions in composing. Since its inception, keystroke logging has been widely used in writing research in the contexts of timed writing tasks (Almond, Deane, Quinlan, Wagner, & Sydorenko, 2012; Barkaoui, 2016; Sinharay, Zhang, & Deane, 2019), academic writing (Choi, 2016), professional writing (Leijten, Van Waes, Schriver, & Hayes, 2013), and language translation (Dragsted & Carl, 2013). There has also been a trend to apply keystroke logging techniques in educational settings to support writing assessment (e.g., Deane, 2014; Deane, Quinlan, & Kostin, 2011; Sinharay, et al., 2019) and writing pedagogy (e.g., Chukharev-Hudilainen, 2019; Lindgren, Sullivan, & Stevenson, 2008).

As an observational tool in writing research, keystroke logging is a relatively new methodological approach. One of the earliest attempts in using keystroke logging to electronically record writers' keyboard typing activities for research purposes was made in Bridwell and Duin (1985). The application of keystroke logging in writing research gained momentum as computer-based writing becomes increasingly prevalent. To date, a number of keystroke logging programs have been developed, among which Scriptlog (Strömqvist et al., 2006), and Inputlog (Leijten & Van Waes, 2013) are widely used in current writing studies.

These programs were designed with different focuses. Specifically, Scriptlog was devised for use in experimental psycholinguistic research while Inputlog aimed for use in studying multimodal professional writing environments. 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. Time stamps for these keyboard and mouse operations are also reported to indicate when the events occur and how long they last. In some advanced keystroke logging programs such as Inputlog, text position information can also be recorded, which allows for finer-grained analyses of writing behaviors such as pauses at the boundaries of different linguistic units and revisions.

To illustrate, Figure 2.2 provides an example logfile generated with Inputlog. The logfile records the writer's keyboard and mouse activities in detail while writing using MS Word. The *Output* column presents a list of readable actions occurred in the process (e.g., "h" means the character "h" is typed in, "SPACE" refers to the spacebar, and "LEFT" refers to a left click on the mouse; see https://www.inputlog.net/ for more detailed explanations of the input actions), and the *Event Type* column indicates the type of these actions (e.g., mouse, keyboard, focus). The time stamps (in clock time and milliseconds) for each input action are also stored in the output. The start time and end time indicate the duration of the actions. Note that for keystroke activities, start and end time refer to the timing of key press and release respectively. In addition, text position information (in *Position* column) is also registered to help keep track of the location of the leading edge while the lengths of both the process text (in *Character Production* column) and the product text (in *DocLength* column) are represented as accumulated numbers of characters typed in. The output file also reports pause time at different locations (e.g., before words, within words, before sentences) in the *PauseLocation* column, allowing for fine-grained pause analyses.

#Id	Event Type	Output	Position	DocLength	Character Production	StartTime	StartClock	EndTime	EndClock	ActionTime	PauseTime	PauseLocation	X	Y
0	focus	Inputlog 7.0.0.2			0	859	00:00:00	859	00:00:00	0	0	TRANSITION		
1	mouse	Movement			0	859	00:00:00	2234	00:00:02	1375	0	INITIAL	1100	489
2	focus	Wordlog.docx - Word			0	10093	00:00:10	10093	00:00:10	0	0	TRANSITION		
3	keyboard	CAPS LOCK	0	1	1	10093	00:00:10	10231	00:00:10	138	0	INITIAL		
4	keyboard	T	0	1	1	10340	00:00:10	10428	00:00:10	88	247	BEFORE SENTENCES		
5	keyboard	CAPS LOCK	1	2	2	10491	00:00:10	10610	00:00:10	119	151	UNKNOWN		
6	keyboard	h	1	2	2	10582	00:00:10	10677	00:00:10	95	91	BEFORE WORDS		
7	keyboard	i	2	3	3	10746	00:00:10	10816	00:00:10	70	164	WITHIN WORDS		
8	keyboard	s	3	4	4	10860	00:00:10	10991	00:00:10	131	114	WITHIN WORDS		
9	keyboard	SPACE	4	5	5	11030	00:00:11	11119	00:00:11	89	170	AFTER WORDS		
10	keyboard	i	5	6	6	11414	00:00:11	11515	00:00:11	101	384	BEFORE WORDS		
11	keyboard	s	6	7	7	11554	00:00:11	11674	00:00:11	120	140	WITHIN WORDS		
12	keyboard	SPACE	7	8	8	11751	00:00:11	11853	00:00:11	102	197	AFTER WORDS		
13	mouse	Movement			8	13187	00:00:13	13765	00:00:13	578	1436	MOUSE	901	665
14	keyboard	c	8	9	9	16010	00:00:16	16099	00:00:16	89	2245	BEFORE WORDS		
15	keyboard	a	9	10	10	16187	00:00:16	16305	00:00:16	118	177	WITHIN WORDS		
16	keyboard	t	10	11	11	16350	00:00:16	16431	00:00:16	81	163	WITHIN WORDS		
17	keyboard	BACK	11	12	12	18216	00:00:18	18306	00:00:18	90	1866	REVISION		
18	keyboard	BACK	10	11	12	18405	00:00:18	18494	00:00:18	89	189	REVISION		
19	keyboard	BACK	9	10	12	18673	00:00:18	18748	00:00:18	75	268	REVISION		
20	keyboard	d	8	9	12	20311	00:00:20	20442	00:00:20	131	1638	WITHIN WORDS		
21	keyboard	o	9	10	13	20455	00:00:20	20544	00:00:20	89	144	WITHIN WORDS		
22	keyboard	g	10	11	14	20596	00:00:20	20684	00:00:20	88	141	WITHIN WORDS		
23	keyboard	.	11	12	15	21605	00:00:21	21714	00:00:21	109	1009	AFTER SENTENCES		
24	keyboard	LEFT	12	13	16	22662	00:00:22	22743	00:00:22	81	1057	UNKNOWN		
26	keyboard	LEFT	11	13	16	22852	00:00:22	22919	00:00:22	67	190	UNKNOWN		
28	keyboard	LEFT	10	13	16	23013	00:00:23	23114	00:00:23	101	161	UNKNOWN		
30	keyboard	LEFT	9	13	16	23313	00:00:23	23412	00:00:23	99	300	UNKNOWN		
32	keyboard	a	8	13	16	24151	00:00:24	24324	00:00:24	173	838	BEFORE WORDS		
33	keyboard	SPACE	9	14	17	24685	00:00:24	24808	00:00:24	123	534	AFTER SENTENCES		
34	mouse	Movement			17	27875	00:00:27	32281	00:00:32	4406	3190	MOUSE	857	1053

Figure 2.2 The Output of a Logfile Generated by Inputlog.

2.2.2 Keystroke logging measures

The logfiles of keystroke activities with their time stamps generated by keystroke logging programs as illustrated in the above example make it possible to quantify a myriad of temporal features of the writing process. In general, some main keystroke logging features commonly used in writing research largely fall into the following categories: pause, revision, burst, and process variance. These are discussed below.

2.2.2.1 Pause

The way keystroke logging records keyboard activities enables us to generate fine-grained measures with high resolution for pause behaviors. With the accurate timing of each keystroke, inter-keystroke intervals (IKI) can be automatically calculated from the log files. IKI is defined as the gap time between two consecutive key presses typically expressed in milliseconds (Chukharev-Hudilainen, Saricaoglu, Torrance, & Feng, 2019). To illustrate, suppose a writer types a character "A" at time 1 and then a character "B" at time 2. One can obtain the IKI between the two characters simply using the formula: IKI = Time2 - Time1.

Due to the motor constraints of human typing, not all IKIs are used to describe pauses. In writing studies, it is a common practice to use a particular IKI threshold to define pauses. According to Limpo and Alvès (2017), there are two kinds of thresholds commonly used in prior research: short pauses with IKIs of somewhere between thirty milliseconds and around two seconds, and long pauses with IKIs over two seconds. In practice, the choice of an IKI threshold is usually contingent on specific research goals in writing studies (Van Waes & Leijten, 2015). For instance, several studies have used the two-second criterion (e.g., Spelman Miller, 2000;

Sullivan & Lindgren, 2002; Wengelin, 2001). Van Waes and Schellens (2003), however, employed a threshold of three seconds to describe pause, which allowed for a clear distinction between different pausing behavior patterns among adult writers.

Writing research investigating pause behaviors using keystroke logging oftentimes draws upon the duration and frequency of pauses calculated from multiple dimensions (e.g., Dich & Pedersen, 2013; Medimorec & Risko, 2016; Van Waes & Schellens, 2003; Van Waes, van Weijen, & Leijten, 2014). These general measures including the number of pauses (in total or per minute), the proportion of pause time (as a % of total writing time), and the average pause length (commonly operationalized as the mean duration of all pauses in text production), have been established as valuable temporal features in writing process. In addition, researchers have also been interested in examining how the durations and frequencies of pause may vary at different locations (e.g., within words, between words, between sentences). Previous studies (e.g., Baaijen, Galbraith, & De Glopper, 2012; Chukharev-Hudilainen et al., 2019; Spelman Miller, 2000) have documented that pause locations are associated with different patterns of pause behaviors which might provide insights into the underlying cognitive processes in writing (the cognitive insights of pause behaviors are thoroughly discussed in the next section).

2.2.2.2 Revision

Revisions are operations of deletions or insertions in writing (Spelman Miller, Lindgren, & Sullivan, 2008). A deletion is defined as the removal of any stretch of characters from a text (Perrin, 2019) whereas an insertion refers to a sequence of activities to add characters to a growing text (except the end). Text production is a complex and recursive process and often teems with various revision activities (Lindgren & Sullivan, 2006). Keystroke logging keeps tracks of online changes made throughout the text production process, such as the text deleted, inserted, and replaced, the time stamps and locations of the revision events, number of revisions, and number of characters before and after revision, all of which supports analyses of different types of online revisions.

In writing research, the general features of revision behaviors are usually examined by calculating the total number of deletions and insertions (Spelman Miller et al., 2008) or the number of such operations per minute (Van Waes & Leijten, 2015) or by dividing the total number of characters/words in the final product using the total number of characters/words produced in the writing process (Van Waes & Leijten, 2015). With cursor position information registered by keystroke logging, some fine-grained revision measures can be obtained. For example, researchers can distinguish between revisions based on their locations, such as revisions at the point of inscription (i.e., at the current end of a text being produced) or after the text has been transcribed (i.e., at a previous point in the text) (Stevenson, Schoonen, & de Glopper, 2006). Lindgren and Sullivan (2006) also referred to the revisions at the point of inscription as pre-contextual revisions as opposed to revisions made after a full textual context has been externalized. Capitalizing on cursor position information in keystroke logs, the distances of revisions can also be obtained by calculating how far the cursor is moved to make the revisions. Thorson (2000), for instance, measured the distances between the position of the flashing cursor and the revision point and distinguished between immediate revisions (distance equal to zero) and distant revisions (distance larger than zero). In addition, keystroke logging also records the text information being changed in writing process, which allows for distinctions of revisions based on the writers' different intentions in writing. According to Stevenson et al. (2006), there are content revisions that alter the information content of the text (e.g., "She is a teacher" is changed to "She is a student"), language revisions that changes only the surface linguistic features of the text by correcting errors in spelling, grammar, and punctuation, or by rewording the linguistic expression without altering the substance of the information (e.g., "She is not happy" is changed to "She is unhappy"), and typographic revisions that refer to the writers' correction of an error caused by accidentally pressing the wrong key(s).

2.2.2.3 Burst

With keystroke logging data, researchers are able to pinpoint and examine the periods in text production in which stretches of texts were continuously produced with no pauses and/or revisions. The fluent production of written language in temporal segments as such was referred to as "burst" (Kaufer, Hayes, & Flower, 1986). Contrary to the disfluent aspects of text production characterized by pauses and revisions, language bursts are fundamentally related to the writer's productivity in text generation. Hence, burst-related features have been widely used as important measures of writing fluency in research (see e.g., Alvès, Castro, & Olive, 2008; Spelman Miller et al., 2008; Van Waes & Leijten, 2015).

Two types of bursts have been distinguished in writing research: P-bursts that refer to the written segments terminated by pauses, and R-bursts that describe the segments terminated by an evaluation, revision, or other grammatical discontinuity (Chenoweth & Hayes, 2001). Research investigating bursts behaviors using keystroke logging usually calculates the frequencies or distributions of P-bursts and R-bursts, or the lengths of these bursts as measured in milliseconds or in characters (Van Waes & Leijten, 2015).

2.2.2.4 Process variance

The time-stamped keystroke activities registered with keystroke logging can also reveal the variability of the writing process. Process variance attends to the dynamics of the writing process in relation to time and thus represents how the writer's fluency may differ at different stages (Van Waes & Leijten, 2015). Process variance is generally measured by first dividing the whole writing process into a certain number of equal time intervals (e.g., 5 or 10) and then calculating the total number of characters produced in the intervals (often normalized to the average number of characters per minute), or to make it more comparable among writers, the proportion of characters produced per interval. The standard deviation of characters produced per interval is also calculated from keystroke logs as an indicator of process variance. In Van Waes and Leijten (2015), two approaches were used to control for the effect of personal typing skills of a writer to facilitate the description of both intra and interpersonal variability in the dynamics of the writing process. One approach is to normalize the absolute variability measure (e.g., the proportion of characters produced per interval by a particular writer) using the so-called *theoretical maximum* which is usually a reference speed (e.g., 400 characters per minute) for an experienced typist during a copy task. The other approach utilizes a *task maximum* which is based on the writer's highest writing speed in the writing task at hand.

2.3 Keystroke Logging and Cognitive Processes in Written Argumentation

2.3.1 Cognitive models of the writing process

From a cognitive perspective, the writing process is characterized with "a complex set of hierarchically arranged cognitive activities, involving problem-solving and decision-making within clearly defined goals" (Kellogg, 1994, p.13). The cognitive approach to writing lays the theoretical basis of keystroke logging with the fundamental rationale that writing process features derived from keystroke logging data can provide inferences about the underlying cognitive processes (MacArthur, Graham, & Fitzgerald, 2008).

Systematic research on the cognitive processes involved in text production gained momentum when Hayes and Flower (1981) began investigating the mental processes college writers employed using protocol analysis. In Hayes and Flower's 1981 model, three cognitive processes were identified in writing composition: *Planning*, *Translating*, and *Reviewing*. Specifically, planning concerns what to say and how to say it. It is further broken down into three mental operations, namely, setting goals, generating ideas, and organizing ideas into a writing plan. Translating converts plans into written text. Reviewing involves rereading and editing to improve the existing text. The coordination of these three processes is assumed to be governed by a control process referred to as the *Monitor*, and accordingly different configurations of the monitor are associated with different writing strategies in composing. Furthermore, Hayes and Flower's model also suggests that the three cognitive processes can be employed flexibly in writing and any subprocess can interrupt or incorporate any other subprocesses in text production. To illustrate, a writer might generate new ideas to compose while reviewing the formulated text thus, simultaneously operating planning and reviewing. On the other hand, planning or reviewing might interrupt translating if a writer stops midway in her writing to plan for the next idea or simply reread the existing text.

The seminal model laid out by Hayes and Flower (1981) served as a catalyst for most of the subsequent research on the cognitive fabric of writing and the architecture of the writing process (Alamargot & Chanquoy, 2001). Chenoweth and Hayes (2003) presented a revised and fully developed version of the writing process model which distinguishes between four subprocesses: The *Proposer*, the *Translator*, the *Transcriber*, and the *Evaluator/Reviser* (see Figure 2.3 for a sketch of the model). The proposer generates ideas for expression via reflection on the plan and on the linguistic or nonlinguistic information related to the plan. In essence, the proposer creates the so-called idea package to be formulated in language. The translator converts the message in the idea package into word strings by resorting to linguistic processes that include lexical access and syntactic frame construction. The transcriber is responsible for turning the word strings into written texts through motor skills. Finally, the evaluator/reviser monitors and evaluates both proposed language that has been transcribed or not yet been transcribed, and makes changes when any problems and inadequacies are identified. These subprocesses of text production are assumed to form a recursive pattern in which the different processes not only occur more or less simultaneously but also interact with each other during text production, leaving irregular traces in the writing process in the form of pauses, movements, insertions, and deletions (Spelman Miller et al., 2008).

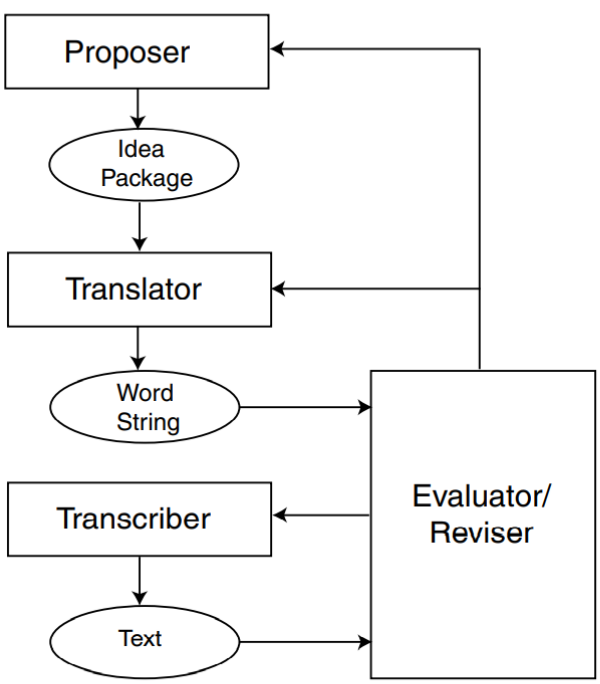

Figure 2.3 The Text Production Model in Chenoweth and Hayes (2003).

In a nutshell, cognitive models of text production demonstrate that skilled writing is a conscious, demanding, and self-directed activity, involving the coordination of a variety of mental operations (MacArthur & Graham, 2016). In this process, the writer constantly capitalizes on their mental resources such as long-term memory, working memory, and attention in order to achieve the desired communication goals (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 writing schemas, mental resources, and cognitive activities in text production 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, Alves, & Castor, 2009; Spelman Miller, 2000; Van Waes et al., 2014).

2.3.2 Aligning keystroke logging measures with cognitive processes in writing

The models of text production as presented above make it possible to interpret the dynamics of the writing process recorded via keystroke logging technologies from a cognitive angle. Whereas alignment between keystroke logging measures and cognitive models of writing has never been void of disputes due to its ambiguous nature (e.g., Baaijen, et al., 2012; Galbraith & Baaijen, 2019; Wengelin et al., 2019), this methodology has arguably provided important pathways in exploring writing behaviors, cognitive efforts and patterns, writing strategies, and their interactions in composing, and hence has become a popular approach in writing research using keystroke logging. Some commonly used keystroke logging measures and their cognitive interpretations are discussed below.

Foremost, pause behaviors as represented with IKIs in keystroke logging are considered as important indicators of writer's covert cognitive activities that are otherwise impossible to observe via merely the written products, such as retrieving a word, rereading pre-transcribed text, and planning for content generation. Pause lengths (e.g., above or below 2 seconds) are assumed to reflect cognitive activities of different levels. According to a general standard set by Limpo and Alvès (2017), pauses between 30 milliseconds and 2 seconds seem to reflect transcription processes (typographic skills and spelling) whereas pauses longer than 2 seconds likely reflect higher order cognitive processes, such as planning for new ideas or revising. In addition, pauses occurring at the boundaries of different linguistic units (characters, words, or sentences) are also assumed to reflect planning or decision-making processes at different textual domains (morphological, grammatical, or discourse levels). Previous research (e.g., Spelman Miller, 2000; Wengelin, 2002, 2006) has generally reported that both pause durations and frequencies increase as text production moves from smaller unit locations (e.g., between words) to larger unit locations (e.g., between sentences). 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 with deleting and inserting activities in keystroke logs, has been acknowledged as an integral cognitive component in the writing process models. It is assumed to be associated with the discrepancies between the writer's intentions and the text generated so far (Leijten, Van Waes, & Ransdell, 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 of 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 assumption underlying this distinction is that global content-related revisions are generally more cognitively demanding than local revisions on surface linguistic features and typographic forms (Stevenson et al., 2006). Revisions at different text positions are also believed to involve different cognitive activities. Lindgren and Sullivan (2006) suggested that pre-contextual revisions (i.e., revisions made at the point of inscription before a full textual context has been externalized) are closely associated with both planning and formulation processes, functioning differently from revisions undertaken after careful re-reading of the previously produced text.

The length of bursts in text production is primarily perceived as an indicator of the writer's cognitive capacity in execution processes (Breuer, 2019). However, bursts delimited by pauses (P-bursts) and by revisions (R-bursts) are associated with different cognitive processes given that they reflect different aspects in text production: P-bursts generally represent the quantity of material that can be formulated in one planning episode while R-bursts reflect output that has been interrupted prior to completion (Galbraith & Baaijen, 2019; Hayes, 2009). According to Galbraith and Baaijen (2019), P-bursts with a pause threshold of two seconds generally relates to the capacity of the translator in text production. The rationale for this assumption is that the two-second pauses represent conceptual planning episodes that generate ideas for the translator. The translator then steadily converts the ideas into linguistic strings till it runs out of content and needs new items of the content. 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.

Instead of pertaining to some specific cognitive subprocesses, the keystroke features of process variance are assumed to reflect how the writer orchestrates these subprocesses during text production, and thus provide insights into the writer's use of strategies in composing (Van Weijen, van den Bergh, Rijlaarsdam, & Sanders, 2009). Although process variance has been rarely explored in writing community, research to date has suggested that more experienced writers generally show a higher degree of process variance in text production often characterized with a distinct pre-writing planning stage and extensive revisions at content and text levels (Van den Bergh & Rijlaarsdam, 2001; Van Waes & Leijten, 2015).

2.4 Keystroke Features and Argumentative Writing Quality

The increasing use of computers in writing tests presents the opportunity to use keystroke logging techniques to capture a stream of information about writing behaviors in a nonintrusive way. Using the temporal features derived from keystroke logging information to characterize the writing process, some writing researchers have been probing into the relationship between these keystroke features and writing performance in assessment contexts.

A flurry of studies in recent years have been conducted to understand and evaluate the links between writing process features and writing performance in the K-12 assessment context and have delineated a set of keystroke features that are predictive of writing quality. For instance, writing time and the number of keystrokes, which are measures of general writing fluency and efforts, have been found related to writing quality with young writers (Zhang, Hao, Li, & Deane, 2016). Another important process feature that has drawn considerable attention from researchers is pause. Deane (2014) found that higher performing students in writing assessment demonstrated shorter pauses at character, word, and sentence boundaries. Likewise, a study conducted by Zhang et al. (2017) suggested that shorter pre-writing pauses under a certain timed-writing test condition indicated adequate understanding of the task requirements, more 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. 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., Dean, 2014).

Process features as those derived from keystroke logging data 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 of their study indicated that better 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 also been reported in research conducted in adult ESL/EFL writing assessment context. 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 predicting 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 what was reported in 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 relationships 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 disputable issue in relating keystroke features to writing quality pertains to 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 that writers' keystroke logging features in text production vary when they respond to tasks of different complexity (Jung, 2017) or different writing genres (Olive, Favart, Beauvais, & Beauvais, 2009). Process features characterized by keystroke activities were also found to correlate with writers' gender (Guo, Zhang, Deane, & Bennett, 2019; Zhu, Zhang, & Deane, 2019), ethnicity (Guo, et al, 2019), working memory capacity (Ransdell, Arecco, & Levy, 2001), language proficiency (Barkaoui, 2019; Van Waes, & Leijten, 2015; Zhu, et al., 2019), among other things. Several studies have attempted to investigate the stability of keystroke logging measures in predicting writing quality across different assessment occasions. Deane and Zhang (2015) investigated the feasibility of modeling writing quality using the keystroke logs collected from a large group of adolescent students in the U.S. responding to six writing test forms that covered two genres (argumentative essay and written recommendation), each containing three different topics. The results suggested that although the stability of keystroke features vary considerably across different writing test occasions, moderate to strong prediction of human-rated essay scores can still be attained with a set of stable features generalizable across writing genres and prompts. 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 showed that despite a large variance in terms of both within-response and within-person stability, most of the keystroke features were found stable and these stable features exhibited meaningful correlations with writing quality. These studies, although limited in their scope and scale, provide promising plausibility in using patterns of keystroke features for measurement of fundamental text production skills as a supplement of human or automated scores on the final product.

2.5 Research Questions

Previous research has shown the value of investigating the process features of text production to better comprehend the cognitive activities involved in producing the final text. An examination of writers' processes in formulating different argument elements would provide valuable pathways to better understand the interplay between writing behaviors, cognition, and writing outcomes in written argumentation. The goal of this dissertation was to examine the links between L1 and L2 writers' behavioral patterns in writing processes manifested by keystroke analytics, the formulation of argument elements, and the holistic quality of their argumentative writing. In this dissertation, L1 participants were U.S. adult residents or citizens who grew up speaking English as the mother tongue while L2 participants were U.S. adult residents or citizens who grew up speaking a language other than English as the primary language. This recruitment standard excluded L2 writers who were EFL writers. The data used in this dissertation contains a set of persuasive essays written by L1 and L2 adult writers with rich keystroke logging information by essay. The keystroke logging measures employed to characterize the adult writers' writing behaviors in text production include bursts, pauses, revisions, and process variances. The argument categories used to describe the argument structures of these persuasive essays are based on Toulmin's (1958, 2003) model of argumentation and comprise final claim, primary claim, counterclaim, rebuttal, and data. Additionally, all essays were scored by trained raters for overall writing quality using a standardized, six-point grading scale. The research questions are as follows:

1. Do keystroke features predict the different categories of argument elements in L1 and L2 adult writers' argumentative writing?

2. Do L1 and L2 adult writers differ in their keystroke features when constructing different categories of argument elements in their argumentative writing?

3. Are there relationships between keystroke features and holistic writing quality in L1 and L2 adult writers?

3 METHOD

3.1 Participants

The data includes four hundred participants (200 L1 and 200 L2 speakers of English) who were hired from Amazon Mechanical Turk (MTurk), an online crowdsourcing platform that enables large-scale participant recruitment in a time-efficient manner. Participants were required to meet three threshold qualifications: 1) be at least 18 years old; 2) be currently living in the United States; 3) have completed at least 50 MTurk tasks with an overall approval rate of at least 98% by experimenters on the platform. The mean age of the participants was 38.11(SD = 11.576). Breakdowns of demographic information for both L1 and L2 participants concerning their gender, age, ethnicity, educational and linguistic backgrounds are presented in Table 3.1.

Table 3.1 Demographic Information for L1 and L2 Participants

Items	L1 Participants (*n* = 200)	L2 Participants (*n* = 200)
Gender	Female (102); Male (94); Non-binary/third gender (3); Prefer not to say (1)	Female (108); Male (90); Prefer not to say (2)
Age	18-30 (51); 31-40 (75); 40-50 (44); 50-60 (23); 61+ (15)	18-30 (70); 31-40 (64); 40-50 (39); 50-60 (30); 61+ (7)
Ethnicity	White (161); Asian (17); Black or African American (8); Hispanic or Latino (7); Multiple ethnicity (5); American Indian or Alaska Native (2)	White (86); Asian (57); Hispanic or Latino (34); Black or African American (14); American Indian or Alaska Native (7); Multiple ethnicity (2)
Education (highest degree completed)	Less than high school diploma (7); High school diploma or the equivalent (15); Some college, but no degree (33); Associate degree (11); Bachelor's degree (98); Master's Degree (32); Professional Degree (2); Doctorate (2)	Less than high school diploma (4); High school diploma or the equivalent (9); Some college, but no degree (19); Associate degree (6); Bachelor's degree (96); Master's Degree (56); Professional Degree (4); Doctorate (6)
L1	English (200)	Spanish (72); French (18); Tamil (15); Chinese (14); Hindi (10); German (7); Japanese (4); Portuguese (4); Russian (4); Swahili (4); Tagalog (4); Arabic (3); Armenian (3); Bengali (3); Urdu (3); Greek (2); Korean (2); Nepalese (2); Romanian (2); Turkish (2); Other (22)
Age of initial English learning		0-5 (74); 6-10 (89); 11-20 (33); 21+ (4)
Years of English learning		0-5 (18); 6-10 (45); 11-15 (62); 16-20 (36); 21+ (39)
Years in the U.S.		0-10 (45); 11-20 (50); 21-30 (51); 31-40 (28); 41+ (26)

3.2 Procedure

Participants hired from MTurk were invited to log onto a website built specially for this study. The website housed a demographic survey, a series of typing tests, an argumentative writing task, and a vocabulary knowledge test. Participants were required to use only computers with a keyboard to participate in the study. Their keystroke activities during the typing tests and the argumentative writing task were recorded using a built-in keystroke logging program that captured every keystroke and mouse operation entered by the participants with its time stamp and cursor position information. Participants' role in the study lasted 40-50 minutes. Their participation was rewarded with a $0.25 reward and a $11.75 bonus upon completing all the tasks following the instructions.

3.3 Materials

3.3.1 Demographic survey

The demographic survey was administrated through Qualtrics Research Suite software (Qualtrics, Provo, UT). It included a set of questions about participants' demographic information concerning their gender, age, citizenship, race/ethnicity, education, etc. For L2 participants, it asked some additional questions about their native languages and English learning experiences (i.e., the age of initial English learning, years of learning English, and years of living in the United States). See Appendix A for a full list of the questions used in the demographic survey.

3.3.2 Typing tests

Given that typing skills are considered an important factor affecting online text production (Van Waes, Leijten, Pauwaert, & Van Horenbeeck, 2019), a series of typing tests adapted from the Inputlog copy task (Van Waes, et al., 2019) were used to measure participants' keyboarding skills. The typing tests consisted of four components designed to gauge participants' typing and motor skills in different conditions. The tests included a tapping test measures which measured the fastest motor speed of pressing two keys (e.g., "d" + "k") alternatively in 15 seconds, a sentence copying test which 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 three-word combination copying test which contained four word combinations and measured the speed of copying a set of low and high frequent bigrams (e.g., "five interesting questions"), and a consonant copy test that assessed typing skills in a non-word context (e.g., "tjxggl"). The typing tests took participants around 5 minutes to complete on average. In this study, the average number of characters in the final products typed per minute in the four components were calculated as a global measure to assess the participants' typing skills.

3.3.3 Argumentative writing task

In the argumentative writing task, participants were asked to write an argumentative essay within 30 minutes 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 participant was randomly assigned one prompt. Appendix B presents the four prompts used in the study. Prior to the writing task, participants were given instructions about the integral components in 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 pages also introduced a set of suggestions for the writing task. These included that participants should write an essay of at least 200 words in 3 paragraphs and that they should not use any online or offline reference materials. To make sure participants stayed focused on the task during writing and to track behavior, the writing task page issued warnings whenever the participant was detected 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 3.1.

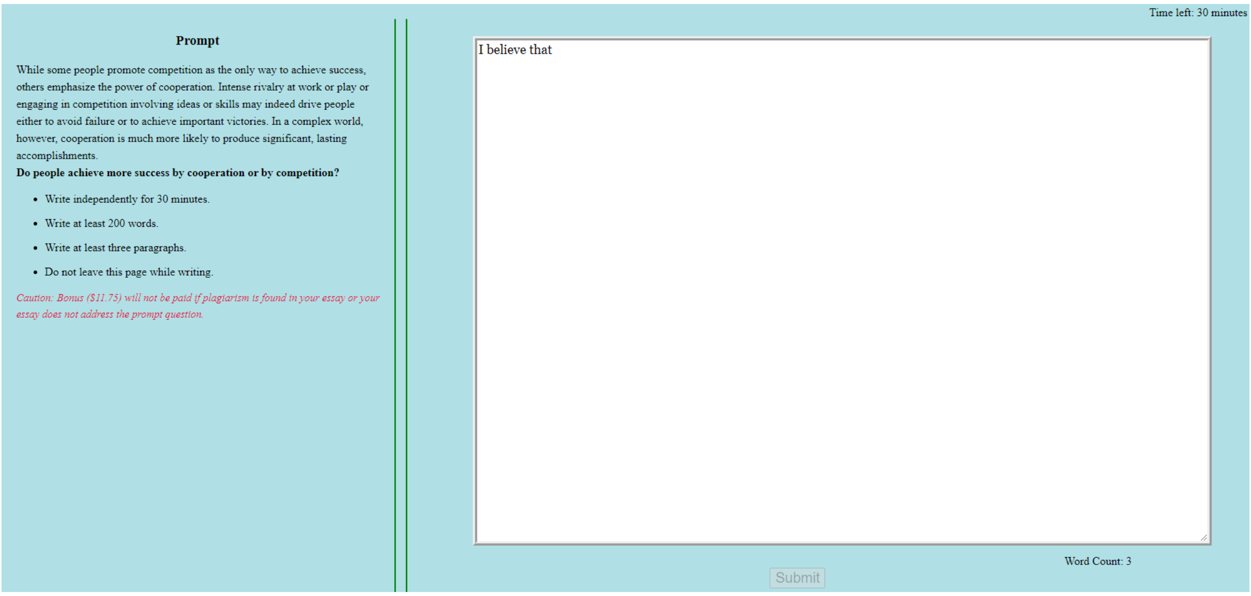

Figure 3.1 A Screenshot of the Writing Task Web Page

3.3.4 Vocabulary knowledge test

Vocabulary knowledge is considered vital in creating meanings during writing (Berninger, 2000; Hayes & Berninger, 2014). To assess participants' vocabulary knowledge, the Lexical Test for Advanced Learners of English (LexTALE) was adapted in the web application. LexTALE was chosen because it is time-efficient and has also been validated as a good predictor of English vocabulary knowledge (Lemhöfer & Broersma, 2011). In this test, participants were presented with 60 trials. In each trial, a string of letters was shown on the screen. Participants were asked to decide whether the string was an existing English word or not by clicking on Yes (word) or No (nonword). A total of 40 English words and 20 nonwords were used for the trials. During each trial, participants were given as much time as needed for their decision. The vocabulary test took around 3-4 minutes to complete on average. The LexTALE scores were calculated using a formula that corrects the unequal proportion of words and nonwords: ((number of words correct / 40 * 100) + (number of nonwords correct / 20 * 100)) /2 (see www.lextale.com).

3.4 Apparatus

To collect participants' keystroke information during the typing tests and the argumentative writing task, a keystroke logging program was written in JavaScript and was embedded in the script of the website built for this study. The program unobtrusively recorded every keystroke operation and mouse activity along with relevant timing and cursor position information when participants completed typing tasks and wrote their essays. It also simultaneously identified operation types (e.g., input, delete, paste, replace) and reported text changes in the writing process. Table 3.2 provides an example output of keystroke logging information reported by the program. *Event ID* indexes the keyboard and mouse operations in chronological order. *Down Time* denotes the time (in milliseconds) when a key or the mouse was pressed while *Up Time* indicates the release time of the event. *Action Time* represents the duration of the operation (i.e., *Up Time* - *Down Time*). *Position* registers cursor position information to help keep track of the location of the leading edge. *Pause Time* demonstrates the time intervals between consecutive events. *Word Count* displays the accumulated numbers of words typed in. Additionally, *Text Change* shows the exact changes made to the current text while *Activity* indicates the nature of the changes (e.g., Input, Remove/Cut).

Unlike some laboratory-based keystroke logging software widely used in writing research (e.g., ScriptLog, Inputlog), this web-based program is more deployable because it automatically runs once its housing webpages are activated and thus avoids inconveniences associated with installing or updating any software beyond a standard web browser. It is also highly scalable and well-suited for online keystroke information collection using crowdsourcing. In terms of privacy concerns, the logging capacity of this program is confined to the input fields on the webpages for the typing tests and the argumentative writing task, and hence does not track of any other information as participants operate on their computers throughout the experiment.

Table 3.2 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

3.5 Essay Annotation

The 400 argumentative essays were human annotated according to the five argument categories specified in Crossley, Tian, and Wan (2022): position, claim, counterclaim, rebuttal, evidence. These elements were originated from Toulmin’s (1958, 2003) model but were modified based on work specific to argumentative writing as found in Liu and Stapleton (2014) and Nussbaum and Kardash (2005). Table 3.3 presents the definitions and examples for the argument elements.

Table 3.3 Definitions and Examples of Argument Elements

Elements	Definitions	Examples
Position	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."
Claim	A claim that supports the position.	"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 statement that refutes the claim or gives an opposing reason to the position.	"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 statement 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."
Evidence	Ideas or examples that support 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."

Three annotators were hired to annotate the argumentative essays using TagTog (https://www.tagtog.net), a web-based text annotation platform. All annotators were graduate students in either Applied Linguistics or English who had experience teaching English writing and coding argumentative texts. Prior to annotating the essays, the annotators were given a workshop on argument categories and the use of TagTog. The annotators were then trained in three norming sessions wherein they independently annotated several batches of argumentative texts from another corpus (36 in total). Their annotation results were then compared and analyzed for differences by the author of this dissertation, who was an expert annotator with over 10 years of experience teaching and researching writing structures and quality. The expert annotator led discussions in the norming sessions and helped resolve disagreements if any. After the norming sessions, the annotators worked independently on TagTog to code the essays randomly assigned to them. Essays were distributed in such a way that each essay was coded by two annotators.

The annotators' independent coding results for the 400 essays were analyzed using Jaccard Index (Tanimoto, 1958) to calculate the degree of overlap between text segments coded by two annotators with the same element category in each essay. 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). In this study, an instance of annotation agreement was determined to occur 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 was 0.806 for all the annotations for L1 essays and 0.866 for L2 essays.

The inter-annotator agreement (IAA) was then calculated for both the L1 and L2 data sets 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 with mutually exclusive categories without any ordering calculations (Bayerl, Lüngen, Gut, & Paul, 2003). The Cohen's Kappa yielded substantial overall agreement between the annotators for both L1 ($k = 0.793$, $p < .001$) and L2 data sets ($k = 0.856$, $p < .001$) according to Landis and Koch (1977).

To calculate how well the annotators agreed with each other when coding for each argument category, the agreement formula found in Ferretti, Lewis, and Andrews-Weckerly (2009) was employed. This formula is Agreement Percentage = Number of Agreement / (Number of Agreement + Number of Disagreement). Table 3.4 presents the agreement percentages for different argument categories coded by the annotators in both L1 and L2 datasets. The results indicate that agreement percentages between annotators were generally high (> 70%) for all argument categories in both L1 and L2 datasets except for the category of claim in the L1 dataset (53.33%).

All the annotations were checked by the third, expert annotator. Wherever there was a disagreement between the 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, the adjudicated annotations were used in analyses.

Table 3.4 Agreement Percentage for Each Argument Category in Both L1 and L2 Datasets

Argument	**Agreement**	
	L1 essays	**L2 essays**
Position (*n* = 379)	89.4%	94.6%
Claim (*n* = 430)	53.3%	72.6%
Counterclaim (*n* = 123)	91.6%	75.7%
Rebuttal (*n* = 116)	74.4%	81.2%
Evidence (*n* = 875)	69%	92.9%

3.6 Essay Quality Rating

The 400 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 C). The holistic rubric evaluates writing quality on multiple dimensions, including test-takers' development of a point of view on the issue, evidence of critical thinking, use of appropriate examples, accurate and adapt use of language, the variety of sentence structures, errors in grammar and mechanics as well as text organization and coherence. Essays were randomly assigned among six raters and each essay was scored by two raters. The raters were graduate students majoring either in 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 400 essays independently. The training started off with a briefing about the essay collection methods and the holistic rubric, followed by a discussion of potential biases in essay scoring. For each session, raters were asked to score a batch of essays on their own before they met to discuss the differences in their scores. A total of 60 practice essays were used for training purposes. These practice essays were on the same topics but were sampled from a different dataset.

A weighted Cohen's Kappa was calculated to measure inter-rater reliability (IRR). The raters exited the training sessions and started independently scoring the 400 essays only after a satisfying agreement had been reached for the practice essays with a Cohen's Kappa of at least .600 (Cohen, 1960). The Initial Cohen's Kappa obtained for the entire dataset after scoring was Kappa = 0.639, $p < .001$, suggesting substantial overall agreement between the raters. If score differences between two raters were two points or greater, the raters adjudicated the scores through discussion. If agreement was not reached, the score was not changed. Among the 400 essays, 31 required adjudication. Average scores between the raters for the adjudicated holistic scores were calculated for each essay and used for the data analysis.

3.7 Keystroke Logging Files Truncating and Segmenting

The first research question asked whether keystroke features predict argument element categories in L1 and L2 adult writers' argumentative writing while the second question investigated if L1 and L2 adult writers differ in their keystroke features when constructing argument elements in their argumentative writing. To address these two questions, the keystroke logging information for each argument element produced in the participants' writing process was obtained and calculated in four main steps as shown in Figure 3.2. First, the keystroke logging information (initially stored in JSON files) was extracted and displayed in tabular CSV dataframes where all keystrokes and their respective time stamps were arranged in a linear order as illustrated in Table 3.2. Second, the text development as it unfolded in the writing process was automatically analyzed by a customized R script written for this dissertation which rearranged and restored the text produced in the process into the product text (the same as it was represented in the submitted essay) based on the keystrokes, time stamps, and cursor positions, and text change information displayed in the dataframe. Third, the restored product text was tagged with argument element labels by comparing the restored text and the argument annotation results. In this way, the timing information for each argument element produced by the participants was obtained. Fourth, the keystroke logging files for each argument element were segmented according to their start and end time information in text production. Note that due to the nonlinearity of the text production process, a participant may not produce text consecutively for a specific argument element. Some participants in this study, for example, were often seen diverting from their effort of building an argument element (e.g., evidence) at the point of inscription to revise or complete another element (e.g., 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 participants. To fix the problem, the segmented log files were merged into the same CSV file for the same argument element. In this merging process, timestamps in the log files that came later were adjusted to avoid exceedingly long pauses at the breaking point using the writers' average pause lengths at specific locations. For instance, if the breaking point occurred between sentences, an averaged between-sentence pause length was applied for the writer.

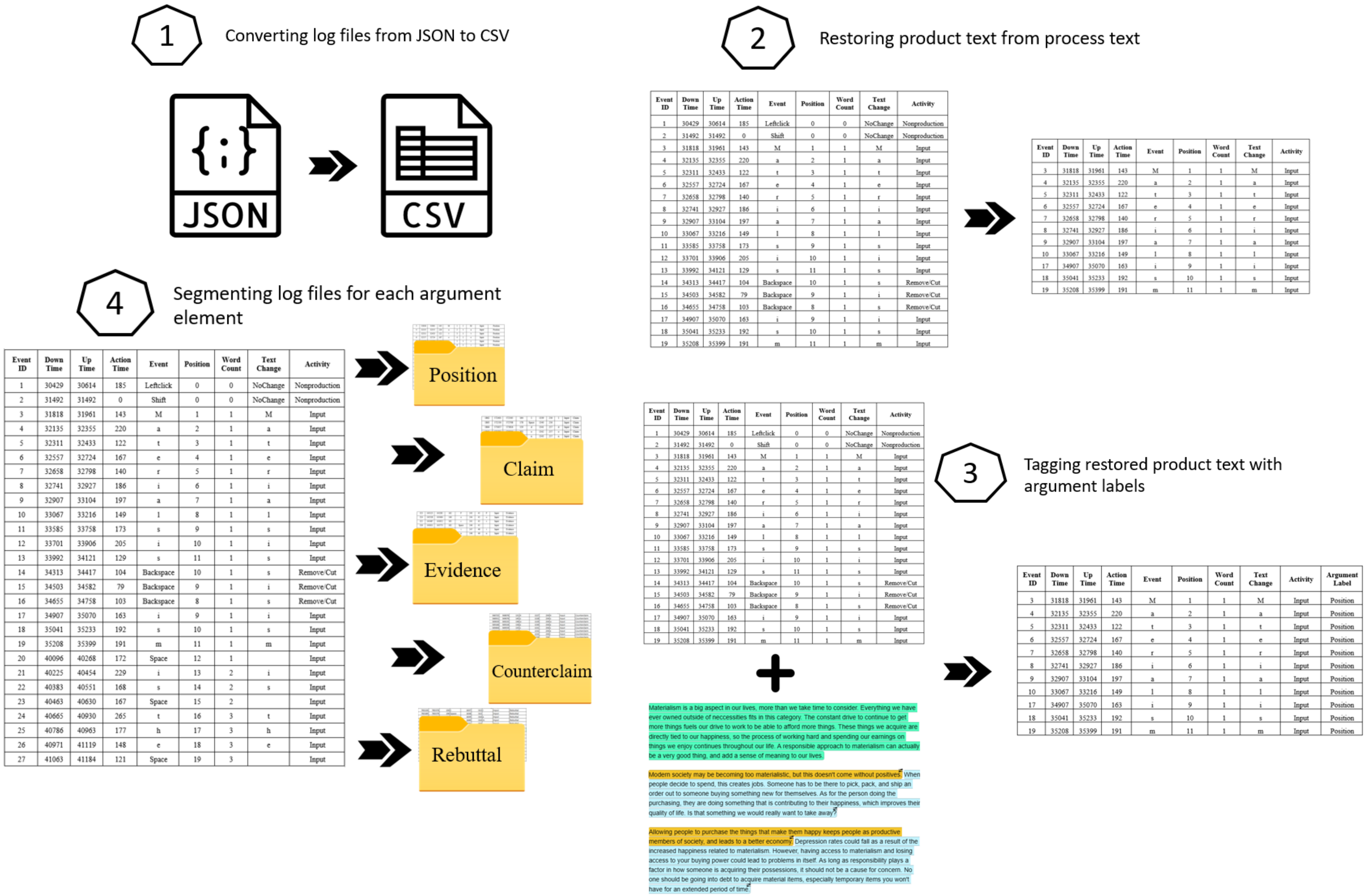

Figure 3.2 Obtaining Keystroke Information for Each Argument Element

To address the third research question which examined if there were relationships between keystroke features and holistic writing quality in L1 and L2 adult writers, the JSON keystroke files for entire formulation processes during the persuasive writing task were first converted into CSV files. They were then checked for any inactive periods at the beginning (e.g., reading instructions, pre-planning) and the end of the processes (e.g., proof-reading, waiting for submission). These inactive periods were removed from the keystroke logging files to delimit the analyses on 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).

3.8 Keystroke Logging Measures

Both the segmented keystroke logging files (mapped to different argument elements) and the truncated files (reflecting entire text production processes) were transformed into IDFX files that could be read and analyzed via Inputlog 7.0 (Leijten & Van Waes, 2015). A set of keystroke indices with reference to the participants' bursts, pausing behaviors, revision activities, and process variances were generated using Inputlog's built-in analysis modules. These were discussed in detail below.

3.8.1 Bursts

Measures of P-bursts were obtained by drawing on the summary analysis provided by Inputlog. In this study, P-bursts were operationalized as continuous text production episodes terminated at pauses of 2 or more seconds following the rationale that pauses exceeding 2 seconds generally corresponding to higher-level thinking processes such as activities related to planning, and thus eliminating the interruptions caused by transcription-related activities or inefficiencies in motor execution. To obtain measures of R-bursts, the revision analysis was performed to calculate the numbers and lengths of bursts delimited by revision activities.

3.8.2 Pauses

Pausing behaviors were analyzed using Inputlog's pause analysis. The threshold for pauses in this study was set 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, 2016). Pauses at different locations (i.e., within words, between words) were also analyzed because they are associated with various patterns of pausing behaviors and might provide insights into different underlying cognitive processes in writing (Chukharev-Hudilainen et al., 2019; Spelman Miller, 2000).

3.8.3 Revisions

Revisions were analyzed by drawing on the revision matrix reported by Inputlog. The revision matrix provides a detailed list of revision events in sequential order. Inputlog distinguishes between two types of revision events: 1) deletions, i.e., characters deleted in the text produced so far, and 2) insertions, which represents the characters inserted into the earlier text. In this study, revision behaviors were operationalized as the proportion of deletions and insertions in the entire text production and the mean length of these two revision types. The former was calculated by dividing the total number of characters produced in the process by the number of characters deleted or inserted. The latter was obtained by calculating the average number of characters for each deletion or insertion.

3.8.4 Process variances

To take into account the variability of each participant's entire writing process, variances in process fluency were investigated using Inputlog's fluency analysis which reports production rates (characters produced per minute) at different stages in the writing process. In this study, process variances were analyzed with the number of intervals set at 10 (i.e., the entire writing time for each essay was divided evenly into 10 segments). To account for personal writing characteristics of the participants, task maximum was used to serve as an individual text-production measure (Van Waes & Leijten, 2015). Inputlog determines the task maximum for each writer in the current writing task by calculating their highest production rate (the number of characters produced per minute) on the basis of a triple moving average applied to consecutive 10 seconds text production intervals. The production rate for each interval was then normed by dividing the writing speed within the interval (i.e., number of characters produced per minute) using the task maximum. Last, the standard deviation was obtained to describe how dispersed the normed production rate for each interval in relation to the mean value.

To limit the inclusion of variables to those that are generalizable and facilitate the comparison of keystroke activities among participants, only keystroke measures based on means, proportions or ratios were selected for statistical analyses. Note that variables that contained null values (e.g., *mean length of between-sentence pauses*, *mean length of between-paragraph pauses*) were also removed. As a result, a total of 10 keystroke measures were retained. See Table 3.5 for descriptions for these 10 measures.

Table 3.5 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.

3.9 Statistical Analysis

3.9.1 Statistical analysis for research question 1 and 2

The first and second research questions investigated whether keystroke measures predict argument elements in L1 and L2 adult writers' argumentative writing and whether L1 and L2 adult writers differ in their keystroke features when constructing different argument elements. Prior to statistical analyses, the keystroke indices were pruned for two reasons. First, some keystroke indices in the dataset have a single value (e.g., 0) 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, 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. Two keystroke indices (*mean length of insertions* and *proportion of insertions*) showed a low percentage of unique values (< 10%) and a high frequency ratio (> 20) and were thus excluded from the analyses. Second, a series of correlation analyses were conducted among all the keystroke indices, participants' typing test, and vocabulary knowledge test scores for the entire datasets to check for multi-collinearity (see Appendix D for the correlation results). Two keystroke indices (*mean length of deletions* and *proportion of deletions*) were found highly collinear (absolute r > .699). In this case, *proportion of deletions* was selected as the representative index. Note that *interval variance* was not calculated for research questions 1 and 2 because a large proportion of the segmented keystroke logs were too short for fluency variability analysis.

The data were analyzed using multinomial mixed effects logistic regression models within a Bayesian framework using the R package *MCMCglmm* (Hadfield, 2010). Bayesian approach was chosen because it is well suited for complex multi-level modeling and is compatible with advanced model fitting techniques such as Markov Chain Monte Carlo (MCMC; Brooks, 1998). In addition, Bayesian methods don't depend on the assumption of normality to determine credible intervals, which circumvents the issue of skewness commonly seen in keystroke logging data (Galbraith & Baaijen, 2019). An advantage of Bayesian statistics over the traditional frequentist approach is that it allows for probability estimation for specific parameter values given the data in addition to the test of the null hypothesis (Levshina, 2016; McElreath, 2020).

Logistic regression models were built for the entire dataset including both the L1 and L2 data. In these models, the five-level argument element category was entered as the dependent variable. Evidence, position, claim, and counterclaim was selected as the reference category respectively for each model to perform a joint mixed-effects linear regression for the log-odds ratio of all possible pairs of argument element categories. The six keystroke measures were entered as fixed effects. To answer the second research question as to whether L1 and L2 groups differ in keystroke features when formulating argument elements, the binary variable *language* (L1 vs. L2) was entered as a fixed effect and two-way interactions between *language* and each of the six keystroke measures were also included. In addition, *typing speed* and *vocabulary knowledge* were entered as control variables. *Prompt* (a categorical variable that has four values: *Appearance*, *Competition*, *Happiness*, *Materialism*) was entered to control for prompt-based effects that might affect written outcomes in argumentative writing (e.g., Knudson, 1992; Liu & Stapleton, 2014; Zhang, 1987). Last, participants' IDs were added as the random effect which provided each participant with a unique intercept to quantify variation across them.

Bayesian statistics require 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, non-informative priors were used, which allows for an approximation of the posteriors based on the data alone. In Bayesian model fitting, MCMC algorithm directly draws samples from the posteriors to approximate the posterior distribution. To achieve a sample size of 5000 for each model, a total of 250,000 iterations were performed 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 output of the models, both measures were reported.

3.9.2 Statistical analysis for research question 3

The third research question addressed the relationships between keystroke features and holistic writing quality in L1 and L2 adult writers. Foremost, the 10 representative keystroke measures were calculated based on the truncated keystroke logs for the actual text production processes during the writing task. These keystroke measures were checked prior to statistical analyses. First, all measures were examined for the issue of near-zero variance given that these variables are commonly considered to have little predicative power in regression models (Kuhn & Johnson, 2013, p 44). 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 10 keystroke measures had a low percentage of unique values (< 10%) and a high frequency ratio (> 20). Second, a series of correlation analyses were conducted among all the keystroke measures, participants' typing test and vocabulary knowledge test scores, and holistic scores of writing quality (see Appendix E for the correlation results). No pair of keystroke measures was found highly collinear (absolute r > .699).

To investigate whether keystroke measures were predictive of human rated holistic scores for writing quality for L1 and L2 adult writers, a series of linear regression models were built for the entire dataset including L1 and L2 data using *R* (R core Team, 2020) and the CARET package (Kuhn et al., 2020). In these models, the outcome variable of holistic scores for writing quality was entered as the dependent variable. The 10 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; Schoonen et al., 2003; Stevenson et al., 2006). Two-way interactions between language and each of the 10 keystroke measures were also tested respectively 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 have been 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): 1) To start with, 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 rerun without the candidate variable. If more than 25% changes were 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.

Meanwhile, k-fold cross-validation was performed for model training and evaluation. K-fold cross-validation is a resampling procedure used to estimate how the model is expected to perform in general in making predictions on data not used during the training of the model. The "k" in k-fold controls the number of subsets into which data is split. Each subset is used once as a test set while the other subsets are combined to form the training set. Thus, in k-fold cross-validation, k number of models are fitted and evaluated allowing for k estimates of performance. In this study, 10-fold cross-validation was performed because it works well on small datasets like the L1 and L2 datasets reported here. For each combination of predictors in the backward variable elimination process, 10 models were developed and each model was evaluated for each fold providing 10 estimates of a model's performance on the ten different test sets. In each round, the best fitting model was selected based on summary statistics such as root mean squared error (RMSE) and mean absolute error (MAE) between the observed and modeled holistic scores of writing quality.

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 Lindeman, Merenda and Gold's (LMG) method (Lindeman et al., 1980) implemented in the R package "relaimpo" (Grömping, 2006).

4 KEYSTROKE LOGGING AND ARGUMENT ELEMENTS IN L1 AND L2 ADULT WRITERS

4.1 Research Question 1 and 2 Results

Table 4.1 and 4.2 display the descriptive statistics of the keystroke analytics for different argument categories in L1 and L2 data respectively. These include the mean values of the keystroke logging indices and their standard deviations. In general, the argument category of evidence had the largest count in both L1 and L2 datasets followed by claim and position. Counterclaim and rebuttal were generally infrequent. This is especially true with L2 writers who produced 39 counterclaims and 36 rebuttals among the 473 argument elements they formulated in total. Table 4.3 shows the means and standard deviations of the typing test scores and the vocabulary knowledge test scores for both L1 and L2 writers. On average, L1 writers exhibited better keyboarding skills and vocabulary knowledge than L2 writers. Table 4.4 presents the results of the fixed effects and two-way interaction effects between *language* and each keystroke measure reported by the logistic regression models, including their posterior means, the 95% CIs, and the frequentist p-values. The details of these results were summarized and reported below. Note that interaction effects were first reported given that main effects become less important and interpretable when they were also found involved in significant interactions.

Table 4.1 Descriptive Statistics of Keystroke Analytics for Each Argument Category in L1 Dataset

Keystroke Logging Indices	**Position (*n* = 194)**		**Claim (*n* = 222)**		**Evidence (*n* = 469)**		**Counterclaim (*n* = 85)**		**Rebuttal (*n* = 80)**	
	M	**SD**	**M**	**SD**	**M**	**SD**	**M**	**SD**	**M**	**SD**
mean length of P-bursts	45.43	33.05	52.44	67.12	67.12	56.46	56.93	35.97	54.52	35.79
mean length of R-bursts	27.68	18.54	27.97	26.95	26.95	18.21	28.33	17.79	28.9	22.75
mean length of pauses	9.41	13.57	19.26	7.32	7.32	15.16	29.14	28.71	31.49	31.28
mean length of within-word pauses	0.45	0.26	0.43	0.43	0.43	0.23	0.38	0.12	0.63	2.1
mean length of between-word pauses	1.51	1.46	1.37	1.43	1.43	1.23	1.22	1.07	1.23	0.88
proportion of deletions	0.09	0.11	0.07	0.09	0.09	0.07	0.08	0.1	0.08	0.07

Table 4.2 Descriptive Statistics of Keystroke Analytics for Each Argument Category in L2 Dataset

Keystroke Logging Indices	**Position (*n* = 185)**		**Claim (*n* = 208)**		**Evidence (*n* = 406)**		**Counterclaim (*n* = 38)**		**Rebuttal (*n* = 36)**	
	M	**SD**	**M**	**SD**	**M**	**SD**	**M**	**SD**	**M**	**SD**
mean length of P-bursts	33.84	25.4	40.53	31.69	43.58	40.86	43.17	33.38	52.15	39.34
mean length of R-bursts	22.25	14.16	22.68	15.42	24.15	17.53	19.92	11.58	20.2	11.34
mean length of pauses	9.05	12.47	14.52	13.77	7.57	25.37	20.38	17.8	23.9	26.61
mean length of within-word pauses	0.68	2.28	0.46	0.33	0.47	0.36	0.5	0.39	0.41	0.18

mean length of between-word pauses	2.3	5.01	2.41	8.24	1.69	1.09	1.47	1.18	1.62	3.3
proportion of deletions	0.1	0.16	0.07	0.09	0.08	0.09	0.09	0.1	0.08	0.09

Table 4.3 Typing Speed and Vocabulary Knowledge in L1 and L2 Dataset

	L1 (*n* = 200)		**L2 (*n* = 200)**	
	M	**SD**	**M**	**SD**
Typing speed	301.12	85.74	226.95	87.53
Vocabulary knowledge	90.69	11.2	77.39	16.15

Table 4.4 Fixed Effects of the Multinomial Mixed Effects Logistic Regression Model

	Posterior mean	Lower boundary of 95% CI	Upper boundary of 95% CI	p-MCMC
Position vs. Evidence				
Intercept	-0.074	-1.26	1.025	0.912
Language: L2	-0.753	-1.601	0.066	0.083
Prompt: Competition	0.004	-0.398	0.375	0.984
Prompt: Happiness	0.123	-0.272	0.505	0.528
Prompt: Materialism	-0.007	-0.421	0.448	0.976
Mean length of P-bursts	-0.018	-0.025	-0.011	0.000***
Mean length of R-bursts	0.008	-0.003	0.019	0.17
Mean length of pauses	0.033	0.015	0.052	0.000***
Mean length of within-word pauses	-0.117	-0.664	0.405	0.71
Mean length of between-word pauses	-0.103	-0.262	0.05	0.19
Proportion of deletions	-0.451	-2.392	1.627	0.672
Language × Mean length of P-bursts	0.008	-0.002	0.018	0.128
Language × Mean length of R-bursts	-0.01	-0.026	0.008	0.259
Language × Mean length of pauses	-0.014	-0.039	0.012	0.263
Language × Mean length of within-word pauses	0.373	-0.251	0.971	0.231
Language × Mean length of between-word pauses	0.161	-0.012	0.334	0.059
Language × Proportion of deletions	1.837	-0.628	4.621	0.168
Typing speed	0	-0.002	0.002	0.733
Vocabulary knowledge	-0.003	-0.014	0.008	0.562
Claim vs. Evidence				
Intercept	-0.809	-1.999	0.384	0.188
Language: L2	-0.135	-1.076	0.84	0.78
Prompt: Competition	0.586	0.155	0.995	0.003**
Prompt: Happiness	0.165	-0.27	0.561	0.438

Prompt: Materialism	0.388	-0.061	0.831	0.094
Mean length of P-bursts	-0.011	-0.017	-0.006	0.000***
Mean length of R-bursts	0.001	-0.01	0.013	0.804
Mean length of pauses	0.08	0.063	0.097	0.000***
Mean length of within-word pauses	-0.136	-0.696	0.403	0.6584
Mean length of between-word pauses	-0.161	-0.335	0.008	0.051
Proportion of deletions	-2.6	-4.911	-0.11	0.028*
Language × Mean length of P-bursts	0.01	0.002	0.018	0.022*
Language × Mean length of R-bursts	-0.015	-0.034	0.004	0.125
Language × Mean length of pauses	-0.034	-0.057	-0.01	0.002**
Language × Mean length of within-word pauses	-0.122	-0.957	0.64	0.754
Language × Mean length of between-word pauses	0.233	0.054	0.431	0.01*
Language × Proportion of deletions	0.587	-2.672	4.004	0.739
Typing speed	-0.001	-0.003	0.001	0.541
Vocabulary knowledge	0	-0.011	0.012	0.949
Counterclaim vs. Evidence				
Intercept	-5.131	-7.752	-2.632	0.000***
Language: L2	0.126	-1.665	2.132	0.915
Prompt: Competition	0.43	-0.309	1.153	0.234
Prompt: Happiness	0.518	-0.173	1.239	0.15
Prompt: Materialism	0.102	-0.66	0.871	0.802
Mean length of P-bursts	-0.009	-0.017	-0.001	0.02
Mean length of R-bursts	-0.003	-0.019	0.012	0.719
Mean length of pauses	0.094	0.076	0.113	0.000***
Mean length of within-word pauses	-1.14	-2.381	0.005	0.217
Mean length of between-word pauses	-0.335	-0.66	-0.019	0.013*
Proportion of deletions	-1.415	-4.477	1.74	0.376
Language × Mean length of P-bursts	0.005	-0.01	0.021	0.542

Language × Mean length of R-bursts	-0.035	-0.075	0.007	0.084
Language × Mean length of pauses	-0.044	-0.073	-0.019	0.000***
Language × Mean length of within-word pauses	0.736	-0.904	2.321	0.424
Language × Mean length of between-word pauses	0.129	-0.421	0.563	0.587
Language × Proportion of deletions	-0.163	-5.669	5.278	0.982
Typing speed	0.001	-0.003	0.004	0.733
Vocabulary knowledge	0.031	0.006	0.055	0.012*
Rebuttal vs. Evidence				
Intercept	-5.498	-7.936	-3.019	0.000***
Language: L2	0.239	-1.334	1.795	0.781
Prompt: Competition	0.523	-0.277	1.266	0.184
Prompt: Happiness	0.697	-0.114	1.423	0.074
Prompt: Materialism	0.1	-0.771	0.951	0.805
Mean length of P-bursts	-0.009	-0.017	-0.001	0.024*
Mean length of R-bursts	0	-0.016	0.016	0.97
Mean length of pauses	0.097	0.078	0.116	0.000***
Mean length of within-word pauses	0.243	-0.16	0.681	0.222
Mean length of between-word pauses	-0.344	-0.656	-0.036	0.015*
Proportion of deletions	-2.555	-5.971	0.569	0.128
Language × Mean length of P-bursts	0.014	0.001	0.027	0.038*
Language × Mean length of R-bursts	-0.041	-0.08	-0.003	0.032*
Language × Mean length of pauses	-0.042	-0.071	-0.013	0.000***
Language × Mean length of within-word pauses	-0.997	-2.656	0.236	0.108
Language × Mean length of between-word pauses	0.333	-0.024	0.663	0.063
Language × Proportion of deletions	0.247	-6.017	6.394	0.907
Typing speed	0	-0.004	0.003	0.835
Vocabulary knowledge	0.029	0.003	0.057	0.028*

Claim vs. Position				
Intercept	-0.682	-2.043	0.573	0.312
Language: L2	0.567	-0.461	1.55	0.278
Prompt: Competition	0.59	0.134	1.111	0.016*
Prompt: Happiness	0.044	-0.418	0.53	0.872
Prompt: Materialism	0.39	-0.086	0.933	0.118
Mean length of P-bursts	0.006	-0.001	0.013	0.093
Mean length of R-bursts	-0.007	-0.019	0.005	0.264
Mean length of pauses	0.049	0.031	0.067	0.000***
Mean length of within-word pauses	-0.042	-0.655	0.56	0.838
Mean length of between-word pauses	-0.062	-0.267	0.128	0.536
Proportion of deletions	-2.163	-4.775	0.586	0.122
Language × Mean length of P-bursts	0.002	-0.008	0.013	0.71
Language × Mean length of R-bursts	-0.004	-0.024	0.018	0.731
Language × Mean length of pauses	-0.02	-0.044	0.003	0.084
Language × Mean length of within-word pauses	-0.508	-1.403	0.261	0.238
Language × Mean length of between-word pauses	0.075	-0.131	0.275	0.471
Language × Proportion of deletions	-1.184	-4.969	2.511	0.531
Typing speed	-0.001	-0.003	0.001	0.391
Vocabulary knowledge	0.004	-0.01	0.017	0.59
Counterclaim vs. Position				
Intercept	-4.789	-7.267	-2.429	0.000***
Language: L2	0.628	-1.01	2.374	0.487
Prompt: Competition	0.389	-0.378	1.14	0.326
Prompt: Happiness	0.399	-0.346	1.144	0.307
Prompt: Materialism	0.059	-0.762	0.853	0.875
Mean length of P-bursts	0.01	0	0.019	0.045
Mean length of R-bursts	-0.012	-0.029	0.004	0.155

Mean length of pauses	0.064	0.044	0.082	0.000***
Mean length of within-word pauses	-0.962	-2.409	0.457	0.226
Mean length of between-word pauses	-0.181	-0.468	0.082	0.184
Proportion of deletions	-1.008	-4.316	2.434	0.566
Language × Mean length of P-bursts	-0.003	-0.02	0.013	0.757
Language × Mean length of R-bursts	-0.019	-0.061	0.015	0.323
Language × Mean length of pauses	-0.03	-0.057	-0.006	0.012*
Language × Mean length of within-word pauses	0.395	-1.378	1.946	0.613
Language × Mean length of between-word pauses	-0.003	-0.364	0.384	0.961
Language × Proportion of deletions	-1.458	-7.049	3.86	0.638
Typing speed	0	-0.003	0.004	0.824
Vocabulary knowledge	0.03	0.007	0.054	0.009**
Rebuttal vs. Position				
Intercept	-5.559	-8.678	-2.846	0.000***
Language: L2	1.831	-0.229	3.765	0.09
Prompt: Competition	0.418	-0.401	1.249	0.324
Prompt: Happiness	0.57	-0.214	1.335	0.15
Prompt: Materialism	0.08	-0.842	0.945	0.846
Mean length of P-bursts	0.008	-0.004	0.02	0.184
Mean length of R-bursts	-0.009	-0.025	0.006	0.292
Mean length of pauses	0.067	0.048	0.088	0.000***
Mean length of within-word pauses	0.376	-0.136	1.004	0.127
Mean length of between-word pauses	-0.289	-0.7	0.106	0.122
Proportion of deletions	-2.146	-5.684	1.395	0.219
Language × Mean length of P-bursts	0.007	-0.009	0.022	0.43
Language × Mean length of R-bursts	-0.041	-0.084	0.004	0.076
Language × Mean length of pauses	-0.028	-0.053	-0.002	0.032*

Language × Mean length of within-word pauses	-3.547	-5.36	-0.613	0.000***
Language × Mean length of between-word pauses	0.231	-0.184	0.647	0.243
Language × Proportion of deletions	-1.645	-8.354	4.973	0.656
Typing speed	-0.001	-0.005	0.003	0.67
Vocabulary knowledge	0.036	0.007	0.063	0.014*
Counterclaim vs. Claim				
Intercept	-3.922	-6.756	-1.218	0.003**
Language: L2	-0.109	-1.849	1.634	0.913
Prompt: Competition	-0.174	-0.933	0.691	0.657
Prompt: Happiness	0.361	-0.509	1.118	0.386
Prompt: Materialism	-0.275	-1.115	0.625	0.509
Mean length of P-bursts	0.003	-0.005	0.012	0.524
Mean length of R-bursts	-0.005	-0.021	0.014	0.562
Mean length of pauses	0.015	0.003	0.025	0.006**
Mean length of within-word pauses	-1.152	-3.765	0.473	0.184
Mean length of between-word pauses	-0.185	-0.501	0.122	0.225
Proportion of deletions	1.078	-2.035	4.412	0.509
Language × Mean length of P-bursts	-0.005	-0.021	0.009	0.521
Language × Mean length of R-bursts	-0.018	-0.06	0.024	0.433
Language × Mean length of pauses	-0.011	-0.028	0.007	0.221
Language × Mean length of within-word pauses	1.353	-0.377	3.853	0.112
Language × Mean length of between-word pauses	-0.038	-0.391	0.352	0.845
Language × Proportion of deletions	-0.389	-6.16	4.961	0.89
Typing speed	0.001	-0.002	0.005	0.518
Vocabulary knowledge	0.027	0.002	0.054	0.037*
Rebuttal vs. Claim				
Intercept	-4.995	-7.589	-2.201	0.000***

Language: L2	0.64	-1.142	2.368	0.48
Prompt: Competition	-0.157	-1.083	0.699	0.737
Prompt: Happiness	0.462	-0.366	1.252	0.26
Prompt: Materialism	-0.316	-1.231	0.524	0.486
Mean length of P-bursts	0.002	-0.008	0.011	0.764
Mean length of R-bursts	-0.001	-0.019	0.014	0.979
Mean length of pauses	0.018	0.008	0.03	0.000***
Mean length of within-word pauses	0.378	-0.041	0.889	0.07
Mean length of between-word pauses	-0.207	-0.561	0.097	0.23
Proportion of deletions	0.425	-3.339	4.356	0.828
Language × Mean length of P-bursts	0.005	-0.008	0.019	0.478
Language × Mean length of R-bursts	-0.029	-0.07	0.013	0.178
Language × Mean length of pauses	-0.009	-0.025	0.01	0.313
Language × Mean length of within-word pauses	-1.53	-2.843	-0.102	0.021*
Language × Mean length of between-word pauses	0.12	-0.246	0.487	0.52
Language × Proportion of deletions	-0.759	-6.993	5.216	0.794
Typing speed	0	-0.004	0.004	0.997
Vocabulary knowledge	0.033	0.007	0.062	0.017*
Rebuttal vs. Counterclaim				
Intercept	-0.99	-3.849	1.833	0.478
Language: L2	0.739	-1.23	2.655	0.521
Prompt: Competition	-0.084	-0.997	0.836	0.854
Prompt: Happiness	0.066	-0.74	0.889	0.91
Prompt: Materialism	-0.024	-0.958	0.95	0.936
Mean length of P-bursts	0	-0.011	0.012	0.988
Mean length of R-bursts	0.003	-0.015	0.021	0.778
Mean length of pauses	0.005	-0.006	0.014	0.37
Mean length of within-word pauses	1.799	0.319	3.353	0.000***

Mean length of between-word pauses	-0.045	-0.453	0.331	0.814
Proportion of deletions	-1.058	-4.736	2.785	0.580
Language × Mean length of P-bursts	0.009	-0.008	0.027	0.327
Language × Mean length of R-bursts	-0.014	-0.059	0.033	0.605
Language × Mean length of pauses	0.001	-0.014	0.017	0.88
Language × Mean length of within-word pauses	-2.892	-4.82	-0.903	0.000***
Language × Mean length of between-word pauses	0.149	-0.297	0.634	0.523
Language × Proportion of deletions	-0.069	-5.818	5.69	0.965
Typing speed	-0.001	-0.005	0.003	0.532
Vocabulary knowledge	0.003	-0.025	0.029	0.799

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

4.1.1 Position vs. evidence

The model found no significant two-way interactions between *language* and any of the keystroke measures in distinguishing positions from evidence. However, *mean length of P-burst* and *mean length of pauses* were identified as significant main effects. The 95% CIs suggest that the odds of participants formulating positions rather than evidence increased when they engaged in shorter P-bursts (i.e., P-bursts with fewer characters) and paused longer on average in text production. None of the other keystroke logging indices significantly predicted positions in comparison to evidence. There were no significant effects for *prompt*, *typing speed*, or *vocabulary knowledge* in predicting positions either.

4.1.2 Claim vs. evidence

In predicting claims against evidence, significant interaction effects were detected between *language* and three keystroke measures respectively, including *mean length of P-bursts*, *mean length of pauses*, and *mean length of between-word pauses.* As illustrated in Figure 4.1, L1 and L2 participants tended to compose shorter P-bursts when formulating claims than when formulating evidence for those claims. However, compared to their L2 counterparts, L1 participants exhibited a steeper drop in P-burst lengths when composing claims than evidence. Figure 4.2 shows that both L1 and L2 participants likely produced text with longer pauses when constructing claims. In comparison to the L2 group, however, L1 participants displayed a more evident increase in pause lengths when formulating claims than evidence. As for between-word pauses, L1 and L2 participants demonstrated distinct patterns (see Figure 4.3). Whereas L1 participants produced text with slightly shorter pauses between words when formulating claims than evidence, L2 participants paused remarkably longer between words in claim construction in comparison to evidence formulation. In addition, *mean length of P-burst*, *mean length of pauses*, and *proportions of deletions* were identified as three significant main effects in predicting claims versus evidence. Among these three main effects, only *proportions of deletions* was not significantly moderated by *language*. Its 95% CIs indicate that the odds of participants formulating claims as opposed to evidence increased when they deleted less. A significant prompt effect was also detected in claim formulation. Specifically, the 95% CIs show that participants likely produced more claims when writing on *Competition* than *Appearance*. No significant effects were found for *typing speed*, or *vocabulary knowledge* in distinguishing claims from evidence.

Figure 4.1 Interaction Effect between Language and Mean Length of P-burst in Formulating Claims Versus Evidence

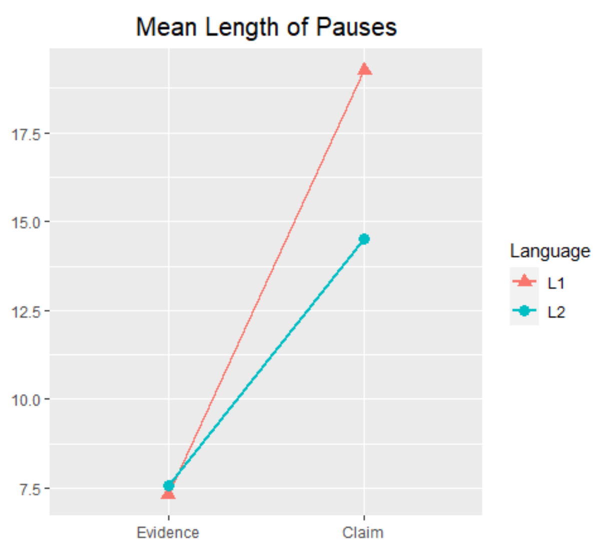

Figure 4.2 Interaction Effect between Language and Mean Length of Pauses in Formulating Claims Versus Evidence

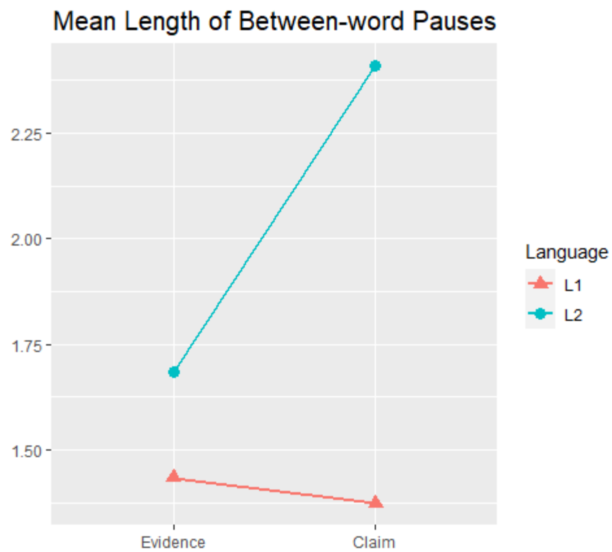

Figure 4.3 Interaction Effect between Language and Mean Length of Between-word Pauses in Formulating Claims Versus Evidence

4.1.3 Counterclaim vs. evidence

In terms of counterclaims versus evidence, the model reported a significant interaction effect between *language* and *mean length of pauses*. As shown in Figure 4.4, both L1 and L2 participants were prone to pause longer when formulating counterclaims. In contrast to the L2 participants, however, a more marked increase in pause lengths were observed in L1 participants' text production for counterclaims than evidence. Apart from the interaction effect, significant main effects were also found for *mean length of pauses* and *mean length of between-word pauses*. Among these two main effects, *mean length of between-word pauses* was not significantly moderated by *language*. The 95% CIs indicate that the odds of participants producing counterclaims instead of evidence increased when they had shorter pauses between words. In addition, *vocabulary knowledge* was also found significant, the 95% CIs of which indicate that participants were more likely to produce counterclaim than positions when they had more advanced vocabulary knowledge. No significant effects were found for *prompt* or *typing speed* in distinguishing counterclaims from evidence.

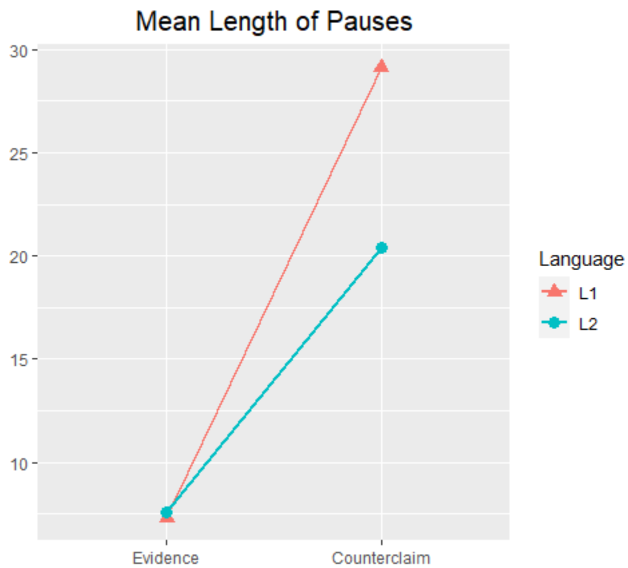

Figure 4.4 Interaction Effect between Language and Mean Length of Pauses in Formulating Counterlaims Versus Evidence

4.1.4 Rebuttal vs. evidence

In predicting rebuttals in comparison to evidence, significant interaction effects were found between *language* and three keystroke measures respectively, namely, *mean length of P-bursts*, *mean length of R-bursts*, and *mean length of pauses*. Results as plotted in Figure 4.5 suggest that L1 participants composed with shorter P-bursts whereas L2 participants wrote with longer P-bursts when they were constructing rebuttals than evidence. This language effect also moderated their R-burst patterns such that L1 participants exhibited longer R-bursts while their L2 counterparts demonstrated shorter R-bursts in producing rebuttals versus evidence (see Figure 4.6). As for pausing behaviors, similar results were found as those in counterclaim formulation (see Figure 4.7). It is shown that both L1 and L2 participants tended to pause longer when formulating rebuttals. Compared to the L2 group, however, L1 participants demonstrated a more dramatic increase in their pause lengths when producing rebuttals than evidence. Furthermore, *mean length of P-bursts*, *mean length of pauses* and *mean length of between-word pauses* were identified as significant main effects. Among these three main effects, only *mean length of between-word pauses* were not significantly moderated by *language*. Its 95% CIs indicate that the odds of participants producing rebuttals rather than evidence increased when they had shorter pauses between words. In addition, *vocabulary knowledge* was also identified significant, the 95% CIs of which suggest that participants were more likely to produce rebuttals instead of evidence when they possessed better vocabulary knowledge. No significant effects were found for *prompt* or *typing speed* in distinguishing rebuttals from evidence.

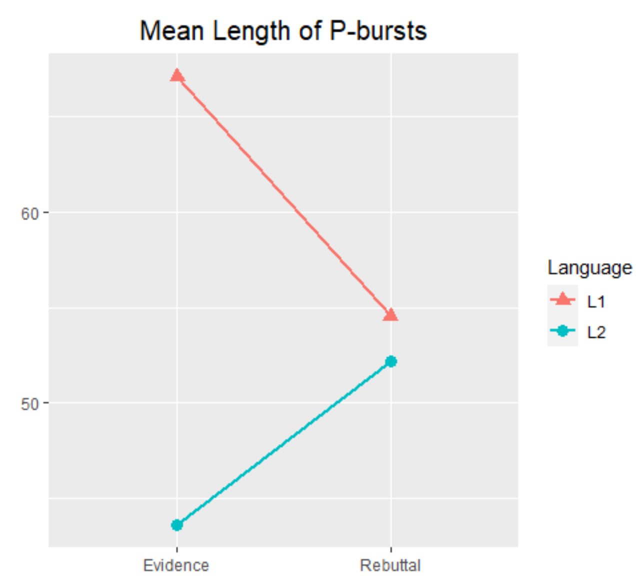

Figure 4.5 Interaction Effect between Language and Mean Length of P-bursts in Formulating Rebuttals Versus Evidence

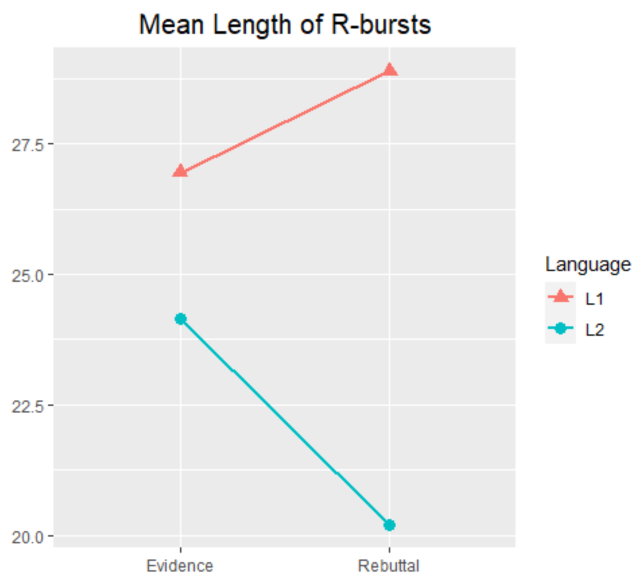

Figure 4.6 Interaction Effect between Language and Mean Length of R-bursts in Formulating Rebuttals Versus Evidence

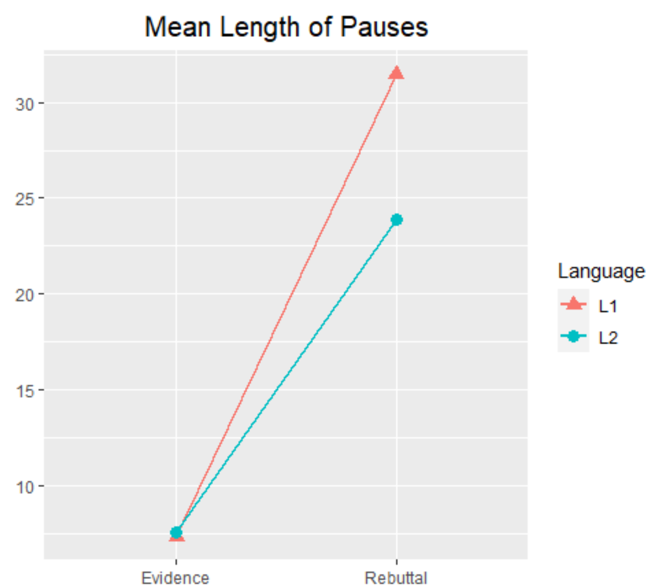

Figure 4.7 Interaction Effect between Language and Mean Length of Pauses in Formulating Rebuttals Versus Evidence

4.1.5 Claim vs. position

In distinguishing claims from positions, the model reported no significant interactions between *language* and any of the keystroke measures. However, the results showed a significant main effect for *mean length of pauses*. The 95% CIs indicate that the odds of participants producing claims in comparison to positions increased when they engaged in longer pauses. In addition, a significant prompt effect was also detected, which indicates that participants likely produced more claims compared to positions when writing on *Competition* than *Appearance.* No significant effects were found for *typing speed*, or *vocabulary knowledge* in distinguishing from claims and positions.

4.1.6 Counterclaim vs. position

In predicting counterclaims versus positions, the model identified a significant interaction effect between *language* and *mean length of pauses*. As demonstrated in Figure 4.8, both writer groups produced text with longer pauses in counterclaim formulation. However, a more evident increase was found for L1 participants in comparison to L2 participants when they were constructing counterclaims than positions. Additionally, *mean length of pauses* was also identified as a significant main effect in the model but there is no need to interpret the results because of its interaction effect with *language*. Moreover, *vocabulary knowledge* was also detected significant in predicting counterclaims in comparison to positions. The 95% CIs suggest that participants were more likely to produce counterclaims as compared to positions when they possessed better vocabulary knowledge. No significant effects were found for *prompt* or *typing speed* in distinguishing counterclaims from positions.

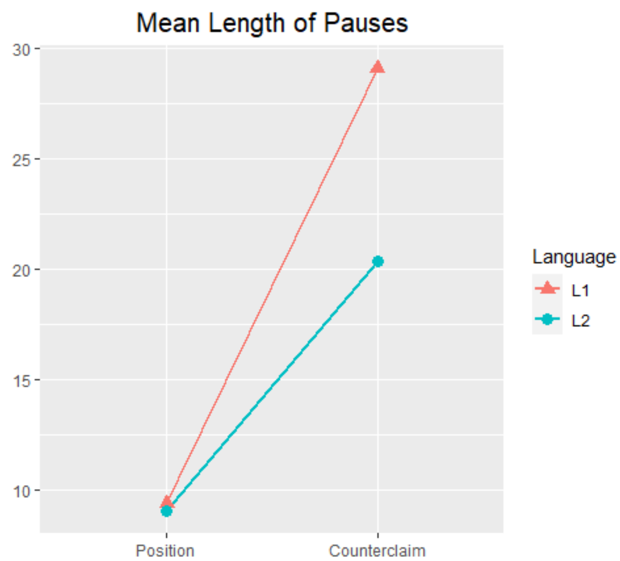

Figure 4.8 Interaction Effect between Language and Mean Length of Pauses in Formulating Counterclaims Versus Positions

4.1.7 Rebuttal vs. position

As for predicting rebuttals in comparison to positions, significant interaction effects were found between *language* and two keystroke measures (*mean length of pauses* and *mean length of within-word pauses*). As shown in Figure 4.9, both L1 and L2 participants exhibited longer pauses when producing rebuttals than positions. However, compared to L2 participants, L1 participants showed a steeper increase in pause lengths when formulating rebuttals versus positions. The two groups also displayed distinct keystroke patterns in terms of pause behaviors within words. Figure 4.10 shows that although L1 participants displayed increased pause lengths within words when formulating rebuttals than positions, the duration of within-word pauses was shorter as L2 participants composed rebuttals than positions. Additionally, significant main effect was also detected for *mean length of pauses* in predicting rebuttals versus positions but the results will not be interpreted because of its significant interaction with *language*. Moreover, *vocabulary knowledge* was also identified as significant in distinguishing rebuttals from positions. The 95% CIs suggest that participants were more likely to produce rebuttals in comparison to positions when they possessed better vocabulary knowledge. No significant effects were found for *prompt* or *typing speed* in distinguishing rebuttals from positions.

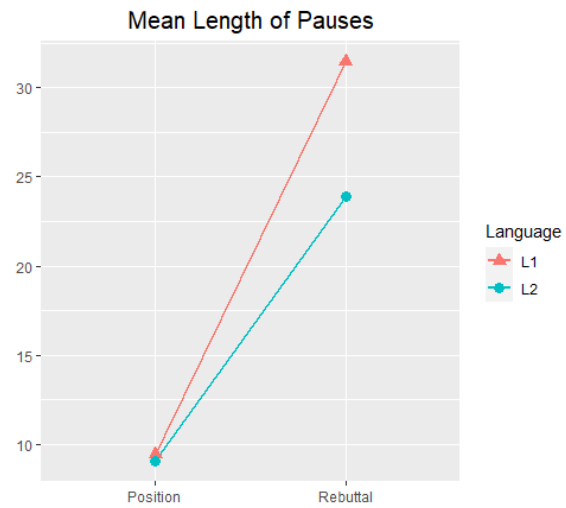

Figure 4.9 Interaction Effect between Language and Mean Length of Pauses in Formulating Rebuttals Versus Positions

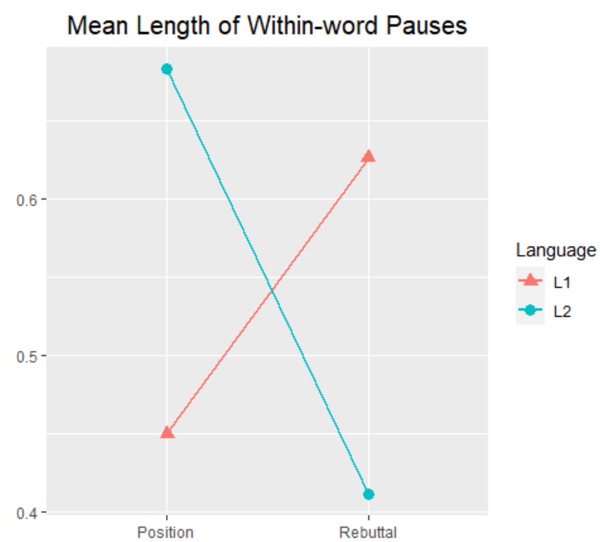

Figure 4.10 Interaction Effect between Language and Mean Length of Within-word Pauses in Formulating Rebuttals Versus Positions

4.1.8 Counterclaim vs. claim

In distinguishing counterclaims from claims, the model identified no significant interaction effects between *language* and any of the keystroke measures. However, a significant main effect was found for *mean length of pauses*. The 95% CIs suggest that the odds of participants producing counterclaims as opposed to claims increased when they composed with longer pauses. In addition, *vocabulary knowledge* was also identified as significant in distinguishing counterclaims from claims. The 95% CIs suggest that participants were more likely to produce counterclaims in comparison to claims when they possessed better vocabulary knowledge. No significant effects were found for *prompt* or *typing speed* in distinguishing counterclaims from claims.

4.1.9 Rebuttal vs. claim

The model distinguishing rebuttals from claims reported a significant interaction effect between *language* and *mean length of within-word pauses*. As illustrated in Figure 4.11, L1 participants paused longer within words when formulating rebuttals in comparison to claims whereas L2 participants exhibited shorter pauses in rebuttal construction than claim formulation. In addition, *mean length of pauses* was identified as a significant main effect. The 95% CIs suggest that the odds of participants producing rebuttals as opposed to claims increased when they had longer pauses in text production. Similar to counterclaim formulation, *vocabulary knowledge* was also found significant in distinguishing rebuttals from claims. The 95% CIs suggest that participants were more likely to produce rebuttals in comparison to claims when they had better vocabulary knowledge. No significant effects were found for *prompt* or *typing speed* in distinguishing rebuttals from claims.

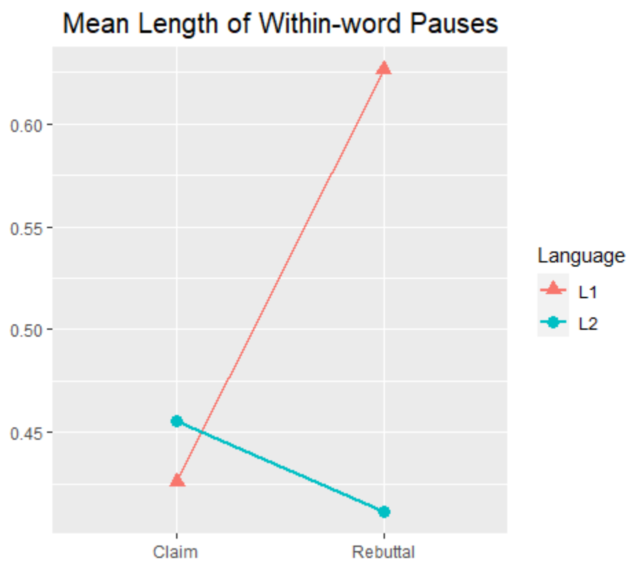

Figure 4.11 Interaction Effect between Language and Mean Length of Within-word Pauses in Formulating Rebuttals Versus Claims

4.1.10 Rebuttal vs. counterclaim

In predicting rebuttals versus counterclaims, a significant interaction effect was detected between *language* and *mean length of within-word pauses*. Results as plotted in Figure 4.12 show different keystroke patterns of pausing behaviors within words between L1 and L2 participants. Whereas L1 participants paused longer within words when constructing rebuttals in comparison to counterclaims, L2 participants exhibited shorter within-word pauses when producing rebuttals than counterclaims. *Mean length of within-word pauses* was also identified as a main significant effect in the model but the results is not interpreted because of the interaction effect already presented here. No significant effects were found for *prompt, typing speed* or *vocabulary knowledge* in distinguishing rebuttals from counterclaims.

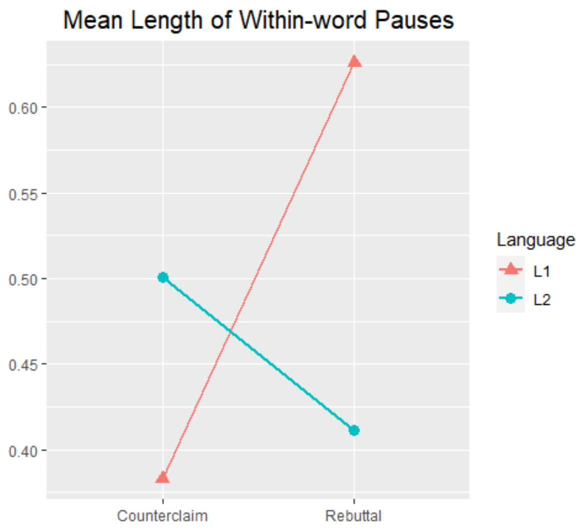

Figure 4.12 Interaction Effect between Language and Mean Length of Within-word Pauses in Formulating Rebuttals Versus Counterclaims

For all the logistic regression models, issues of autocorrelation were checked 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. The Markov chains were also examined as to how well they 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).

4.2 Research Question 1 Discussion

The first research question addressed whether keystroke features of burst, pause, and revision predict the categories of argument elements (i.e., position, claim, evidence, counterclaim, and rebuttal) formulated by L1 and L2 adult writers in argumentative writing. Put together, results reported from the multinomial mixed effects logistic regression models identified a set of keystroke measures that significantly distinguished between the argument element produced by L1 and L2 participants. These include *mean length of P-bursts, mean* *length of pauses*, *mean length of between-word pauses*, and *proportion of deletions*. See Figure 4.13 for a presentation of how values of these four keystroke measures fluctuated when participants were formulating different argument elements. Note that the results regarding *mean length of R-bursts* and *mean length of within-word pauses* are not discussed for the first research question because 1) their effects on argument formulation were highly moderated by *language*, and 2) only contrasting keystroke patterns related to these measures were observed across L1 and L2 participants. Results concerning these two measures are thus merely discussed for research question 2. Overall, these results suggest that both L1 and L2 participants engaged in distinct keystroke patterns in formulating different argument elements. In addition, the logistic regression models also reported a significant prompt effect in claim production. Significant effects were also found for vocabulary knowledge in the formulation of counterclaims and rebuttals. However, no significant effects were found for keyboarding skills. These results are discussed below.

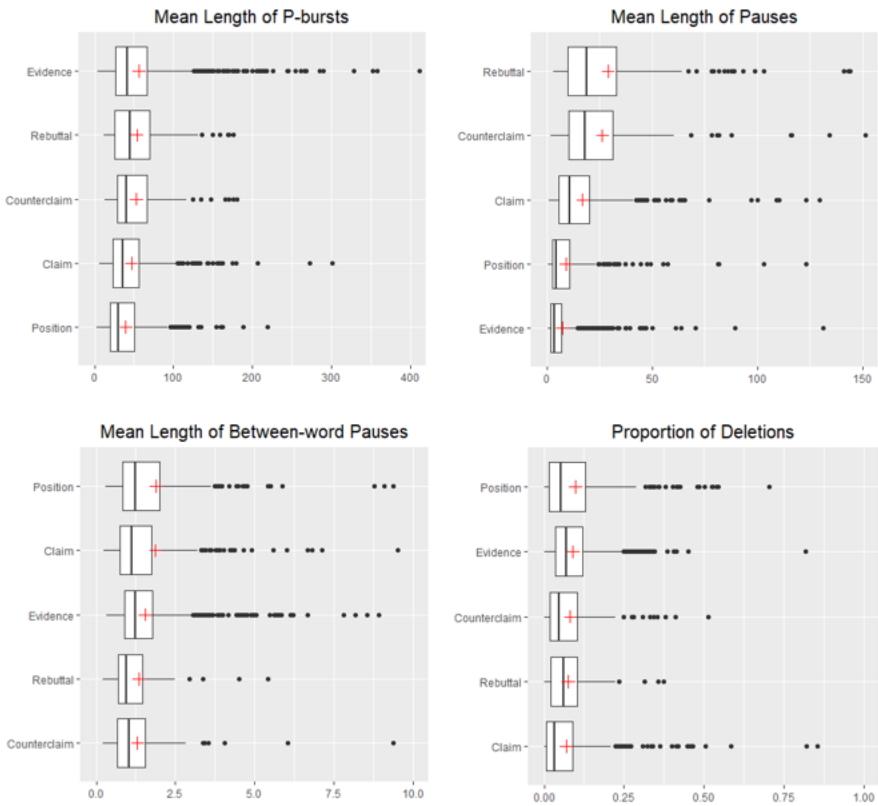

Figure 4.13 Differences in Keystroke Measures When Participants Formulate Different Argument Elements

4.2.1 Keystroke patterns in position formulation

In comparison to the formulation of evidence and rebuttal, the production process of positions featured significantly shorter P-bursts (i.e., smaller numbers of characters produced on average). In terms of pauses in general, positions were constructed with pauses significantly shorter than those in formulating claims, counterclaims, and rebuttals but longer than those found in evidence formulation.

The fewer characters produced within the P-bursts in position formulation suggest that L1 and L2 writers translated their ideas less fluently when constructing positions compared to other elements. Given that the length of P-bursts is commonly recognized as an indicator of individual linguistic capacity (Galbraith & Baaijen, 2019), a plausible explanation is that writers in this study might have undergone more linguistic constraints in producing positions. In comparison to the formulation of other elements, writers might have experienced more rigidity in articulating the overarching position to address the prompt head-on in clear and concise language, as expected in argumentative writing traditions. Another possible contributing factor might be related to the shorter lengths of positions compared to other elements (e.g., evidence) in argumentative discourse, which likely affected the P-burst lengths therein.

Interestingly, the relatively shorter pauses observed reveal another facet of the cognitive process in position production. Compared to the formulation of rebuttals, counterarguments, and claims, positions might not be as cognitively costly to construct. In the argumentative writing task used in this dissertation, it is likely that writers had an initial planning stage prior to their actual composing process when they read their assigned prompts, pondered on the prompt questions, and mentally formulated and rehearsed their stances on the questions. This type of initial planning, which can be done rapidly, implicitly, and unconsciously (Torrance, 2016), could have contributed to better developed ideas and thus led to shorter pauses in general in position production. In addition, the priming effect from the writing prompt might also play a contributing role in reducing the cognitive demand in position formulation. To advance a position in an argumentative essay, the writer needs to propose his/her assertion in response to the contentious topic or problem as described in the writing prompt. In many cases, the writing prompt (as those used in the current study) provides alternative stances or possible solutions as well as essential linguistic cues to prompt the writers in position formulation. For example, the writer may compose a position (e.g., "Yes, I believe modern society should be criticized for being materialistic.") in direct response to the prompt question (e.g., "Should modern society be criticized for being materialistic?").

4.2.2 Keystroke patterns in claim formulation

The formulation of claims was characterized with significantly shorter P-bursts in comparison to evidence construction. In terms of pauses, claims were composed with longer pauses as compared to the formulation of evidence and positions but shorter pauses than that of counterclaims and rebuttals. In addition, the proportion of deletions was significantly lower in claim production than in evidence construction. Note that discussions about the keystroke patterns of between-word pauses in formulating claims versus evidence are left for research question 2 because of the contrasting results for L1 and L2 writer groups.

The P-burst patterns in claim formulation may indicate that writers experienced similar linguistic constraints as they did when formulating positions although the intensity of these linguistic constraints was comparatively lower. To illustrate, writers in this study might try to meet the specific genre expectations of composing terse claims that outline or summarize the overarching ideas for the supporting evidence, which likely cramped their full capacity to translate pre-conceived ideas. Similar to positions, the shorter P-burst lengths in claim production might also be attributed to the intrinsically short length of claims in themselves.

The longer pauses as observed in claim production in comparison to the formulation of evidence and positions may indicate more cognitive efforts involved. In extended discourse such as persuasive writing, the use of claims to address a contentious issue from multiple dimensions signals increased structural complexity in argumentation. Indeed, claims serve as important nodes to connect the overarching position and the supporting evidence, sustaining an unbroken chain of extended arguments (Crammond, 1998). This feature of claims might impose additional demands on writers' working memory in advancing claims that not only echoed the main points embodied in the position but also set the tone and scope for the supporting evidence. Compared to counterclaims and rebuttals, however, claims might be comparatively less cognitively demanding to construct as demonstrated by the shorter durations of pauses in claim formulation. These are further discussed in the sessions for counterclaims and rebuttals.

Additionally, the lower proportion of deletions in claim formulation suggests that writers' claim formulation was a relatively more linear and straightforward cognitive process. One plausible explanation is that claims are oftentimes closely tied to the position they support by restating a certain point in the position but involving specific cases or circumstances linked to the positions (Crammond, 1998), which might help reduce the cognitive load in the transcription process. For example, a writer may simply "borrow" certain words or expressions from previously stated position (e.g., "In my opinion, cooperation is surely the better way to achieve success compared to competition") to the ensuing claims (e.g., "Cooperation is a better way to achieve success because it allows people to work together in harmony.").

4.2.3 Keystroke patterns in evidence formulation

Evidence was produced with significantly longer P-bursts in comparison to positions and claims and shorter pauses in general compared with all the other argument elements. In terms of revision, the formulation process for evidence featured significantly more deletions than that for claim. Note that discussions about the keystroke patterns of P-bursts in formulating evidence versus rebuttals are left for research question 2 because of the contrasting results for L1 and L2 writer groups.

The relatively longer P-bursts in evidence formulation indicate that writers translated their messages into linguistic strings more fluently for evidence than for positions and claims. It is possible that when composing evidence to flesh out or support the general statements in their positions or claims, writers possessed more flexibility and freedom to choose what to include as evidence as long as they are relevant and verifiable. They perhaps also enjoyed more leeway in molding and phrasing the supporting reasons, facts, and details for the evidence without the level of linguistic inhibitions as they might experience in position and claim formulation. The shorter pauses observed in writers' evidence production also point to the relatively less cognitive demand involved as compared to the formulation of other argument elements, which might further substantiate the decreased cognitive constraints in idea generation, content development, and language formulation linked to evidence composing.

However, the larger proportion of deletions indicates that evidence formulation is by no means an easy and linear cognitive process. Indeed, providing relevant and valid evidence to justify the positions and claims entails a set of complex reference, thinking, and reasoning skills (Schwarz & Asterhan, 2010). This is especially true in timed independent writing where the writers may need to critically evaluate and analyze the situation while simultaneously retrieving from their long-term memory relevant knowledge and information, all of which poses demand for attentional resources in writers' working memory.

4.2.4 Keystroke patterns in counterclaim and rebuttal formulation

Excluding the variance associated with the effect of language, L1 and L2 adult writers' keystroke behaviors in the formulation of counterclaims and rebuttals demonstrated quite similar patterns. Significantly longer pauses were observed when writers were composing counterclaims and rebuttals compared with positions, claims, and evidence. However, the durations of between-word pauses in counterclaim and rebuttal formulation were shorter than those observed during the text production for evidence.

The significantly longer pauses detected in the formulation of counterclaims and rebuttals in comparison to other argument elements generally suggest that counterargumentation was a cognitively more costly process for adult writers in written argumentation. The challenges posed by counterargumentation have long been acknowledged in writing scholarship (e.g., Coirier, Andriessen, & Chanquoy, 1999; Nussbaum & Schraw, 2007). Counterargumentation is cognitively demanding because it requires the writer to temporarily identify with an audience with opposing views, imagine its potential objections, and rebut these objections to further strengthen the argument. This process involves increased cognitive efforts in juggling supporting elements, alternative views, and counterevidence along with substantial epistemological sophistication and perspective taking on the part of the writer (Hays & Brandt,1992), all of which likely exerts considerable demands on the writers' limited working memory, leading to interruptions in text production.

However, the shorter duration of between-word pauses observed in counterargumentation in comparison to evidence formulation indicates that writers demonstrated relatively more ease in word retrieval and sentence construction. In this sense, it seems that the cognitive challenges writers encountered in counterargumentation might be tied more to conceptual or pre-linguistic preparation (e.g., conceiving ideas, anticipating audience response) than to language production in itself.

4.2.5 Effects of prompt, keyboarding skills, and vocabulary knowledge in argument formulation

In order to better explain the variances in L1 and L2 adult writers' keystroke behavioral patterns when constructing written arguments, the effects of prompt, keyboarding skills, and vocabulary knowledge were also controlled for in logistic regression modeling. Foremost, results show that there was a significant prompt effect on L1 and L2 writers' formulation of claims in comparison to position production such that the likelihood of writers composing a claim rather than a position increased when they were writing on *Competition* instead of *Appearance*. One possible contributing factor for this prompt effect may be associated with writers' topic knowledge that potentially impacts argumentative writing outcomes (e.g., He & Shi, 2012; Liu & Stapleton, 2018; Tedick, 1990; Yang & Kim, 2020; Yoon, 2021). It is likely that writers in this study might be more familiar with the topic of competition versus cooperation in achieving success (the object of enquiry in the *Competition* prompt) than the effect of appearances on society (the object of enquiry in the *Appearance* prompt), which led to more proposed supporting reasons as claims when writing on the former than the latter. Another plausible explanation is that the differences in rhetorical functions of the two prompts might also played a role in adult writers' critical thinking in written argumentation (Liu & Stapleton, 2018). In comparison to the *Appearance* prompt which required writers to evaluate one single social phenomenon, the *Competition* prompt, which is a typical comparison/contrast writing task, allowed writers to compare and evaluate two different approaches to success achievement, likely prompting writers to respond from at least two perspectives and thus eliciting more claims to back up the position.

Significant effects for vocabulary knowledge were also found on counterargumentation in participants' argumentative writing. Specifically, the results reported in this dissertation suggest that the chance of formulating a counterclaim rather than a piece of evidence or a position increased for writers with better vocabulary knowledge. Likewise, for those who possessed better vocabulary knowledge, the likelihood of constructing a rebuttal instead of evidence also improved in text production. The links between adult writers' use of counterargumentation and their vocabulary knowledge might not be surprising given that vocabulary knowledge has long been established as an important indicator of writing abilities and performances (e.g., Cumming, 1990; Grabe & Kaplan, 2014; Laufer & Nation, 1995). It is assumed that better vocabulary knowledge facilitated word retrieval and transcription, leaving more attentional resources in writers' working memory for higher-level cognitive activities, including envisioning an audience and refuting the opposing views it might hold.

4.3 Research Question 2 Discussion

The second research question investigated whether L1 and L2 adult writers differ in their keystroke features when formulating different argument elements in writing. The results reported from the logistic regression models suggest that the two writer groups demonstrated a set of significantly different keystroke patterns they were 1) producing claims versus evidence, and 2) when they were constructing counterclaims and rebuttals in comparison to other argument elements. These are discussed in detail below.

Foremost, in formulating claims versus evidence, L1 and L2 writers exhibited distinct keystroke patterns in reference to P-bursts, pauses in general, and pauses between words. Specifically, whereas L1 writers generally produced text in longer P-bursts for both claims and evidence, the length of their P-bursts demonstrated a more marked decrease when composing claims than evidence. In terms of the duration of pauses, although L1 and L2 writers both produced text with longer pauses for claims than for evidence, a more evident increase was found for L1 writers. As for between-word pauses, L1 and L2 writers demonstrated contrasting keystroke patterns. While L1 writers' pauses between words slightly decreased when constructing claims versus evidence, L2 writers demonstrated a remarkable increase in their between-word pauses. These results can be interpreted from two perspectives: strategic and linguistic. First, it seems likely that L1 writers underwent a more salient increase in their cognitive efforts when producing claims versus evidence as indicated in their keystroke patterns of P-bursts and pauses. The greater change in the cognitive load on the part of L1 writers might reflect their more conscious allocation of cognitive resources in writing based on a better knowledge of the argumentative genre conventions (e.g., the roles of claims versus evidence). For example, L1 writers might be highly aware of the functions of claims in written argumentation and hence devote more attention to conceiving overarching claims that are relevant, valid, and effective in governing the supporting evidence. Second, these results might also reflect the gap in linguistic skills between the two language groups. The steeper change in the length of P-bursts by L1 writers in formulating claims versus evidence might be largely attributed to the fact that L1 writers produced text with considerably longer P-bursts than L2 writers (see Figure 4.1). This may showcase L1 writers' higher linguistic capacity in translating ideas into written text than their L2 counterparts. Moreover, the drastic increase in between-word pauses during L2 writers' claim formulation indicates their struggles in leveraging linguistic knowledge and resources for word retrieval and sentence formulation.

In constructing counterclaims and rebuttals as compared to evidence and positions, a consistent tendency was detected in L1 and L2 writers' pausing behaviors. Whereas both L1 and L2 writers paused longer in counterargumentation, L1 writers exhibited a greater increase in their pause lengths compared to their L2 peers. Again, these different pausing patterns might be related to how L1 and L2 writers allocated their cognitive resources in written argumentation. The greater increase in L1 writers' pause lengths when formulating counterclaims and rebuttals may reflect their special attention and efforts in considering and evaluating alternative viewpoints and further strengthening their arguments by refuting potential objections. Indeed, essay annotations for this dissertation show that L1 writers were generally more prone to craft counterclaims and rebuttals in written argumentation than L2 writers (see Table 4.1 and 4.2).

A more varied picture was revealed concerning L1 and L2 keystroke pattern differences in rebuttal formulation. Specifically, in composing rebuttals versus evidence, L1 writers demonstrated a decrease in P-burst lengths and an increase in R-burst lengths while L2 writers composed with longer P-bursts and shorter R-bursts. These results suggest that in rebuttal formulation, L1 writers largely experienced more constraints on their linguistic capacity to translate ideas into text (shorter P-bursts) and their continuous text production were less likely interrupted by revision activities (longer R-bursts). In contrast, L2 writers generally underwent fewer constraints on the capacity of the translator (longer P-bursts) and their continuous text production were more likely to be interrupted by revision operations (shorter R-bursts). These results may indicate L2 writers' struggles in adeptly using English to express ideas due to their lower English proficiency. As for within-word pauses in composing rebuttals as compared to all the other argument categories, L1 adult writers produced longer pauses within words whereas L2 writers demonstrated shorter within-word pauses. Put together, these distinct patterns might point to different strategies L1 and L2 writers resorted to in the translation and transcription processes during rebuttal formulation. It may be that L1 writers tended to pause when their linguistic capacity became overloaded in crafting rebuttals versus other elements, leading to shorter P-bursts and longer within-word pauses. On the contrary, L2 writers might revise the text rather than pause when the cognitive demand overwhelmed their translator, resulting in shorter R-bursts and shorter within-word pauses.

5 KEYSTROKE LOGGING AND WRITING QUALITY IN L1 AND L2 ADULT WRITERS

5.1 Research Question 3 Results

Results from the final linear regression model are displayed in Table 5.1. The results show that participants' holistic scores were significantly predicted by typing speed, vocabulary knowledge, language, and 5 keystroke measures, namely, *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 participants with higher essay scores tended to pause less in general and particularly between words. They were less likely to delete what they produced but were prone to engage in more insertions. Their text production processes featured less variance in production fluency. In general, participants who produced higher quality argumentative text generally had better keyboarding skills and vocabulary knowledge. Additionally, L1 participants scored significantly higher than L2 participants. However, no significant interaction effect 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 5 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 were presented in Table 5.2. As shown, *typing speed* and *vocabulary knowledge* were the two most important variables in the model, yielding a total of 75.1% contributions to the R^2 values. This is followed by *language*, which accounted for 1.5% contributions alone. The five keystroke logging measures in total made up 13.3% of contributions in 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 5.1 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 5.2 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%

5.2 Research Question 3 Discussion

The third research question in this dissertation examined the associations between writing process features as manifested by keystroke measures and the holistic text quality of the final products in L1 and L2 adult writers' argumentative writing. A set of keystroke measures were identified as significant predictors of holistic writing quality. These include two measures of pausing behaviors (*mean length of pauses* and *mean length of between-word pauses*), two measures of revising behaviors (*proportion of deletions* and *proportion of insertions*), and one measure of process variance (*interval variance*). Overall, the results verify previous research that the dynamics of the writing process as observed via keystroke logging techniques are predictive of the holistic text quality of the final products (e.g., Allen et al., 2016; Choi & Deane, 2020; Xu, 2018).

The associations between keystroke features of pausing behaviors and human-rated writing quality scores suggest that adult writers who scored higher generally produced text with shorter pauses, which is in line with previous scholarship (e.g., Allen et al., 2016; Choi & Deane, 2020). The overall shorter pause durations on the part of higher performing writers indicate that their text production was smoother with fewer substantial interruptions due to planning for new ideas, re-reading earlier text, or other types of distractions. It is possible that these higher scorers might have allocated and leveraged their time and cognitive resources more strategically in their writing processes so that they were less likely to temporarily suspend writing due to working memory overload or in some extreme cases, divert their attention from what they were currently composing to other irrelevant activities. Of additional interest in the results is the revelation that writers' pausing behaviors between words were closely associated with their writing performance such that higher scored writers generally paused less between words in text production. Between-word pauses are directly representative of variations in content planning (Galbraith & Baaijen, 2019). This suggests that more proficient writers might have experienced a higher degree of automaticity in word retrieval, clause and sentence structuring, and content development. This may also mean that these writers were less overwhelmed in handling the linguistic constraints and cognitive demands when composing argumentative texts, and thus were able to spare more attentional resources to higher level thinking such as generating ideas, anticipating an imagined audience, and monitoring the logical flow in the discourse.

Two dichotomous global keystroke measures of revision (*proportion of deletions* and *proportion of insertions*) were found to be predictive of human-rated writing quality scores. Specifically, the results show that writers who produced higher quality argumentative text likely composed with fewer deletions but more insertions. The distinct roles of deleting and inserting behaviors in contributing to the final text quality indicate that different cognitive operations might be involved when writers made deletions and insertions. Although findings from previous writing scholarship investigating the links between revising behaviors and writing outcomes have been more or less mixed, 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). It is possible that a large proportion of deletions, oftentimes made by writers at the point of inscriptions, were related to low-level, convent- and rule-governed changes such as spelling and grammatical error corrections. In contrast, insertions, which were executed by moving the cursor to the earlier text, were more likely related to changes at higher linguistic domains (word, clause, and sentence levels) associated with meaning altering and content manipulation.

Also of note is the negative relationship between the text production variability during the writing process (operationalized as *interval variance* in this dissertation) and the text quality of the final product. Although Van Waes and Leijten (2015) reported that the text production of more proficient writers (L1 writers) was characterized by more process variance compared to that of less proficient writers (L2 writers), the findings in this dissertation suggest that a high degree of variability in the writing process might not be a positive contributing factor to high quality text. A plausible explanation might be related to the way keystroke logs were trimmed for keystroke information analysis and measurement. In this dissertation, keystroke measures were calculated based on the truncated writing process where the prolonged inactive periods before and after the text production stage were not accounted for due to the unpredictability of these stages in online data collection. Therefore, the measure of process variance in this dissertation may be more directly related to the stability of the formulating stage per se when writers turned the blueprints of their essays into linguistic strings. In this sense, a lower level of process variance might be a reflection of more sustained attention and effort in wrestling with the constellation of cognitive demands in argumentative writing for continued text production.

The regression models also reported significant effects for keyboarding skills and vocabulary knowledge on text quality in adult argumentative writing. The results verify previous research which demonstrated that better keyboarding skills relate to higher performance in writing (e.g., Alves, Castro, & de Sousa, 2007; Barkaoui, 2014) and better vocabulary knowledge also leads to higher writing quality (e.g., Albrechtsen, Haastrup, & Henriksen, 2008; Milton, Wade, & Hopkins 2010). Motor execution has long been recognized as having a cognitive cost in text production (e.g., Bereiter & Scardamalia, 2013; Cooper & Matsuhashi, 1983; Graham & Harris, 2000). It is presumable that writers with poor keyboarding skills might be forced to direct considerable attention and cognitive resources to motor activities, exerting a toll on other processes and components of writing such as planning and revising. In contrast, better keyboard skills might mean less effortful motor execution and thus more attentional resources for higher-level thinking, leading to better writing outcomes. In a similar vein, a rich vocabulary knowledge repertoire might provide writers with more facility in translating ideas into linguistic strings and in turn transcribing the linguistic strings into written text, which helps free up resources in working memory for higher-level cognitive activities in argumentative writing. Additionally, results from the regression models also confirmed the disparity in writing performance between L1 and L2 writers such that L1 writers generally scored higher than their L2 counterparts, which is not surprising given the wealth of research substantiating the language effects in argumentative writing (e.g., Crossley, & McNamara, 2009; Ferris, 1994; Silva, 1993; Stevenson et al., 2006).

Of particular interest is the relative importance of all the predictors discussed above in the final regression model. The results show that the variance in the holistic writing quality scores of writers' argumentative essays were best explained by their keyboarding skills and vocabulary knowledge, followed by their language status. Although the five keystroke features of the writing process were also found predictive of text quality, they only accounted for a relatively small percentage (13.3%) of the total variance. It is important to note that writers' keystroke features in text production may not directly contribute to the writing outcomes because these features might simply reflect the dynamics of the writing process as writers navigate through a variety of cognitive activities to convert their ideas into written text. This, however, does not eclipse the values of keystroke measures which provide important inferences about writers' behaviors, strategies, and the underlying cognitive activities in text production.

6 CONCLUSION AND OUTLOOK

The purpose of this dissertation was to investigate the interplay between behavioral patterns in the writing process as manifested in keystroke activities (i.e., pause, revision, burst, process variance), the formulation of argument elements (i.e., positions, claims, evidence, counterclaims, and rebuttals), and the holistic writing quality in L1 and L2 adult writers. The inquiry was further broken down into three overarching research questions. A summary of the results for these questions is presented below followed by discussions of limitations of this dissertation and future research directions.

6.1 Summary of the Results

The first research question dealt with whether keystroke measures of writing behaviors (i.e., burst, pause, and revision) predict different argument element categories (i.e., position, claim, evidence, counterclaim, and rebuttal) constructed by L1 and L2 adult writers in argumentative writing. The multinomial mixed effects logistic regression models built on the entire dataset reported significant effects for four keystroke measures in distinguishing the argument categories from each other without significant moderation from language effect. These measures include *mean length of P-bursts*, *mean length of pauses*, *mean length of between-word pauses*, and *proportion of deletions*. The results generally suggest that adult writers in this study engaged in distinct keystroke patterns in formulating different argument elements (see Figure 4.13). Specifically, in formulating positions, writers produced text with shorter P-bursts (compared with the formulation of evidence and rebuttals) and shorter pauses (compared with the formulation of claims, counterclaims, and rebuttals). In formulating claims, writers composed with relatively shorter P-bursts (compared with evidence formulation), longer pauses (compared with evidence formulation), and lower proportion of deletions (compared with position formulation). When constructing evidence, writers in this study wrote with shorter pauses (compared with the formulation of positions, claims, counterclaims, and rebuttals). In crafting counterclaims and rebuttals, writers demonstrated longer pauses in general (compared with the formulation of positions, claims, and evidence) but shorter pauses between words (compared with evidence formulation).

The second research question asked whether L1 and L2 adult writers differ in their keystroke features when formulating different argument elements. To answer this question, interaction effects between *language* and each keystroke measure were tested in the logistic regression models and a set of significant interactions were identified. Generally, these results show distinct keystroke patterns between L1 and L2 writers when they produced claims in comparison to evidence and when they formulated counterclaims and rebuttals compared to other argument elements. Specifically, in constructing claims versus evidence, although both groups generally produced shorter P-bursts and longer pauses, L1 writers demonstrated a more evident drop in the length of P-bursts and a more marked increase in the duration of pauses. Moreover, L1 and L2 writers also demonstrated contrasting patterns of between-word pauses when composing claims in comparison to evidence. Whereas L1 writers' between-word pauses slightly decreased, L2 writers demonstrated a considerable increase in their pauses between words. In formulating counterclaims and rebuttals as compared to the construction of other elements, although both writer groups paused longer, L1 writers exhibited a more remarkable increase in their pause lengths compared to L2 writers. In composing rebuttals versus evidence, L1 writers exhibited a decrease in P-burst lengths and an increase in R-burst lengths whereas L2 writers composed with longer P-bursts and shorter R-bursts. Last, in formulating rebuttals versus all the other elements, L1 writers demonstrated longer pauses within words while L2 writers produced text with shorter within-word pauses.

The third research question investigated the associations between the dynamics of the writing process as manifested by keystroke measures and the human-rated holistic scores of text quality in L1 and L2 adult writers' argumentative writing. Even though most of the variance could be ascribed to keyboarding skills, vocabulary knowledge, and to a lesser degree first language, the multiple linear regression models built for the entire dataset identified a set of keystroke measures as significant predictors of holistic writing quality, including two measures of pausing behaviors (*mean length of pauses* and *mean length of between-word pauses*), two measures of revising behaviors (*proportion of deletions* and *proportion of insertions*), and one measure of process variance (*interval variance*). These results suggest that higher writing scores were related to shorter pauses in general, shorter between-word pauses, lower proportion of deletions, higher proportion of insertions, and less process variance. The

6.2 Implications

6.2.1 Theoretical implications

Theoretically, this dissertation contributes to the writing research community in mainly two aspects. First, the results of this dissertation shed light on the interplay between the dynamics of the writing process, argument formulation, and writing outcomes in L1 and L2 adult writers. Hence, this dissertation provides a window into adult written argumentation that includes analyses of both adult writers' argument formulation in the writing process and the argument features in their written products as well as the links between these two facets of argumentative writing. Additionally, the results also contribute to a better understanding of L1 and L2 differences in written argumentation, which is of much value given the paucity of research on written argumentation in current L2 writing scholarship (Hirvela, 2017).

Second, in this dissertation, keystroke logging techniques were employed to unobtrusively observe adult writers' text production in written argumentation. With keystroke logging data, a set of fine-grained process measures in reference to writers' pausing, revising, and language burst were generated to provide detailed information as to how writers behaved in formulating different arguments, what strategies they took to navigate challenges in writing, and how they allocated time throughout the whole process. Moreover, this dissertation also drew on cognitive models of text production to obtain inferences about writers' cognitive activities involved in written argumentation from their keystroke features. This process-based approach to written argumentation shows potential in writing research and will perhaps serve as a catalyst for more studies that use keystroke logging to investigate the associations between the writing process and a variety of other linguistic and rhetorical features in the written products.

6.2.2 Pedagogical implications

The results of this dissertation also afford pedagogical implications. Foremost, the findings reveal that adult writers demonstrate different behavioral patterns when constructing different argument elements, which suggests that the formulation of different categories of argument elements might be associated with different cognitive resources and thus leads to varied degrees of cognitive demand. A clear tendency in the results is the relatively more cognitive cost on the part of adult writers when engaging in counterargumentation in comparison to the construction of other elements. The challenge of crafting counterclaims and rebuttals in written argumentation is also evidenced by the smaller counts of these categories found in the adult writers' essays. Another discernible pattern as revealed in the results is the increased linguistic constraints adult writers experienced when formulating positions and claims. These results corroborate the importance of interventions in argumentative writing instruction aimed to enhance adult learners' ability in counterargumentation as well as language support that familiarizes adult learners with linguistic registers in written argumentation (particularly in formulating positions and claims).

The findings of this dissertation may also inform L2 argumentative writing instruction. Although it seems to be a common pedagogical practice to borrow frameworks and approaches proved effective in L1 context to L2 classrooms (see e.g., Bacha, 2010; Chason, Loyet, Sorenson, & Stoops, 2017; Liu & Stapleton, 2014; Salter-Dvorak, 2016), the results in this dissertation show that the two writer groups demonstrated different writing behaviors and strategies in written argumentation as manifested in their keystroke activities. As for the writing outcomes, L2 writers were less likely to strengthen their argumentation using counterarguments and rebuttals and they generally produced lower-quality persuasive essays compared with their L1 counterparts. All these may suggest the necessity of heuristics developed specially for L2 adult learners in argumentative writing instruction.

Last, the process approach to L1 and L2 adult written argumentation in this dissertation may also speak for the values and potential of using keystroke logging for pedagogical purposes in writing instruction in general. Foremost, keystroke logging may offer a diagnostic technique to help teachers monitor learners' writing development and in vivo processes. With the aid of keystroke logging information, writing instructors may be able to observe learners' writing process such as whether and where they have long pauses, whether they generate text fluently, and where and what they revise. This knowledge can be aggregated to provide teachers with evidence of learners' writing strategies and writing difficulties in text generation, which can serve as points for instructional decision making. To illustrate, writing analytic tools that integrate keystroke logging technology may be able to visualize writers’ text production processes and highlight writing episodes that involve prolonged pauses and frequent revisions. Ideally, these tools may also support the playback of the highlighted episodes to help teachers and students better examine how text production unfolds in time and where interventions are needed. Additionally, keystroke logging information may be used to direct learners' attention to their text production process instead of merely the final products, which likely helps boost learner autonomy, metacognitive awareness, and self-regulation in writing (Spelman Miller, 2005; Van Waes, Leijten, Wengelin, & Lindgren, 2012).

6.2.3 Practical implications

The dissertation also delivers practical implications. By delineating a set of keystroke measures that effectively distinguish between different argument categories and predict human ratings of writing quality in adult writers, this dissertation may help point to promising directions for the development of automated writing evaluation (AWE) and intelligent tutoring systems (ITS). Currently, the majority of AWE and ITS systems rely solely on automated analyses of learners' completed drafts despite the recognition of pedagogical values of process-oriented information in writing research and instruction (Chukharev-Hudilainen, 2019). Integrating keystroke logging into these intelligent writing evaluation and support tools may not only help improve the accuracy of these tools in automated writing assessment but also allow for process-based evaluations and feedback to accommodate the learners' real-time needs in text production. For instance, intelligent writing support tools may use keystroke logging information to better detect writers’ irregular writing behaviors such as protracted pauses (mind-wanderings or difficulties in idea generation) and frequent revisions (struggles in idea formulation or spelling depending on the types of the revisions). Keystroke information like this would enable the tools to provide hints or suggestions that help writers to better navigate difficult stages in the writing process. Furthermore, the links revealed in this dissertation between keystroke measures and argument features in adult argumentative writing also indicate the potential for AWE and ITS systems to use keystroke logging for more accurate and pertinent discourse-level feedback and support in argumentative writing.

6.3 Limitations and Future Research Directions

There are several limitations in this dissertation that need to be acknowledged. First, the counts of counterarguments and rebuttals were smaller compared to other element categories found in participants' essays, especially in the case of L2 writers. The relative paucity of these elements might compromise the generalizability of results and implications with regard to counterargumentation because results from the logistic regression models built in this dissertation might be swayed by the idiosyncratic nature of a small group of audience-aware writers who were adept in crafting counterarguments to address potential objections to their argument. Future studies on written argumentation may proactively handle this problem by giving writers especially L2 writers a clear and detailed goal instruction prior to the writing task to elicit more counterclaims and rebuttals (Nussbaum & Kardash, 2005). If research design and experiment time length permit, some short-term instructional interventions on counterargumentation before administering the writing test can also help to encourage more production of counterarguments (Liu & Stapleton, 2014).

Second, in this dissertation, a timed independent argumentative writing task was administered as a window to observe and analyze writers' text production in written argumentation. Although this writing task resembles argumentative writing in assessment settings such as the writing components in some well-known standardized writing tests (e.g., SAT, TOFEL, and IELTS), it does not represent the full picture of argumentative writing practices, especially those commonly seen on a daily basis in personal, professional, and academic contexts where the writing process features more recursive characteristics with no strict time limits for planning, drafting, and revising activities. Additionally, the writing task used in this dissertation does not reflect the writing activities related to source use which is essential in academic writing (Ferretti & Lewis, 2018). In this sense, the results yielded in this dissertation need to be taken with caution when generalized to written argumentation in general.

Another limitation of this dissertation lies in the scope of its inquires. In this dissertation, the argument features in the L1 and L2 adult writers' persuasive essays were analyzed merely based on annotations of discrete Toulmin argument elements. Nevertheless, the links between the argument elements as well as their relative positions or distances from one another in the text are also important structural features worth investigating (see e.g., Azar, 1999; Crossley et al., 2022; Ferretti et al., 2009). Additionally, the strengths of the argument elements constructed by L1 and L2 writers may be equally important in contributing to the whole picture of adult written argumentation (see e.g., Hughes & Lavery, 2015; Means & Voss, 1996; Stapleton & Wu, 2014). More research endeavors may be needed to probe into the associations between the dynamics of the writing process and relational features of argument elements as well as their strengths in written argumentation.

Last, in this dissertation, inferences about L1 and L2 adult writers' cognitive activities were drawn from their keystroke behaviors based on cognitive models of writing (e.g., Chenoweth & Hayes, 2003) widely used in writing research. These inferences, although crucial in providing pathways to exploring writing behaviors, writing strategies, and cognitive efforts and patterns, should be taken with caution. The alignment of keystroke measures with cognitive processes is not void of disputes due to its indirect and ambiguous nature (e.g., Baaijen, et al., 2012; Galbraith & Baaijen, 2019; Wengelin et al., 2019). To illustrate, pinpointing the location of the pauses (at the boundaries of words, sentences, or paragraphs) and associating the pauses with planning for certain linguistic units is disputable because this practice links pausing only with forward planning while neglecting backward planning which might also be related to goal setting and idea generation (Spelman Miller, 2006). This alignment problem also exists with interpretation of burst behaviors in text production. For example, although P-burst length has generally been assumed to indicate the capacity of the translator or the writer's language capacity, it is observed that the length of P-burst may be strongly affected by the writer's strategy for combining translation and reflection (Baaijen, et al., 2012). Therefore, future research on writers' cognitive efforts in text production may sensibly combine keystroke logging data with other sources of evidence including eye-tracking, think-aloud protocols, or stimulated retrospective interviews to give better interpretations of writers' cognitive activities involved in the writing process.

In sum, this dissertation points to several future research directions that may contribute to an improved understanding of argumentative writing processes. The process-based approach adopted in this dissertation demonstrates the potential of using keystroke logging to observe and investigate a myriad of features in argumentative writing, including but not limit to specific linguistic units or characteristics (e.g., formulaic expressions, organization markers, summary statements), relational structures of arguments (e.g., a claim supported by several pieces of evidence may also back up a higher-level claim), and strengths of arguments. In this dissertation, participants’ keyboarding skills and vocabulary knowledge were also included into the analyses as control variables along with the keystroke measures, which added to the explanatory power of the statistical models. Future research may incorporate more factors that potentially impact the writing process such as writers' individual differences (e.g., demographic characteristics, working memory capacity) and task types (e.g., independent vs. source-based writing).

REFERENCES

Alamargot, D., & Chanquoy, L. (2001). *Through the models of writing: with commentaries by Ronald T. Kellogg & John R. Hayes* (Vol. 9). Springer Science & Business Media.

Albrechtsen, D., Haastrup, K., & Henriksen, B. (2008). *Vocabulary and writing in a first and Second language: Processes and development*. Palgrave Macmillan.

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.

Alvès, 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.

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.

Anderson, R.G., Nguyen-Jahiel, K., McNurlen, B., Archodidou, A., Kim, S., Reznitskaya, A., et al. (2001). The snowball phenome- non: Spread of ways of talking and ways of thinking across groups of children. *Cognition and Instruction*, 19(1), 1-46.

Andrews, R. (1995). *Teaching and learning argument*. London, NY: Cassell.

Applebee, A. N.(1984). Writing and reasoning. Review of Educational Research Winter, 54(4), 577-596.

Azar, M. (1999). Argumentative text as rhetorical structure: An application of rhetorical structure theory. *Argumentation*, *13*, 97-114.

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

Bacha, N. N. (2010). Teaching the academic argument in a university EFL environment. *Journal of English for Academic Purposes*, 9, 229-241.

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.

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.

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.

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.

Bereiter, C., & Scardamalia, M. (1982). From conversation to composition: The role of instruction in a developmental process. *Advances in instructional psychology*, *2*(1-64).

Bereiter, C., & Scardamalia, M. (Eds.). (2013). *The psychology of written composition*. Routledge.

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.

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.

Brick, B., & Holmes, J. (2008, October). Using screen capture software for student feedback: Towards a methodology. In *IADIS international conference on cognition and exploratory learning in the digital age*.

Bridwell, L. S., & Duin, A. (1985). Looking in-depth at writers: Computers as writing medium and research tool. *Writing on-line*, 115-121.

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.

Chan, S. (2017). Using keystroke logging to understand writers' processes on a reading-into-writing test. *Language Testing in Asia*, 7(1), 1-27.

Chandrasegaran, A. (2008). NNS students' arguments in English: Observations in formal and informal contexts. *Journal of Second Language Writing*, 17(4), 237–254.

Chason, L., Loyet, D., Sorenson, L., & Stoops, A. (2017). An Approach for Embedding Critical Thinking in Second Language Paragraph Writing. *TESOL Journal*, *8*(3), 582–612.

Cheng, F. W., & Chen, Y. M. (2009). Taiwanese argumentation skills: Contrastive rhetoric perspective. *Taiwan International ESP Journal*, *1*(1), 23-50.

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

Chenoweth, N. A., & Hayes, J. R. (2003). The inner voice in writing. *Written communication*, *20*(1), 99-118.

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.

Choi, Y. H. (1988). Text structure of Korean speakers' argumentative essays in English. *World Englishes*, *7*(2), 129-137.

Chukharev-Khudilaynen, E. (2014). Pauses in spontaneous written communication: A keystroke logging study. *Journal of Writing Research*, 6(1), 61.

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.

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.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1), 37-46.

Coirier, P., Andriessen, J., & Chanquoy, L. (1999). From planning to translating: The specificity of argumentative writing. *Foundations of argumentative text processing*, 1-28.

Connor, U. (1984). A study of cohesion and coherence in English as a second language students' writing. *Research on Language & Social Interaction*, *17*(3), 301-316.

Connor, U. (1990). Linguistic/rhetorical measures for international persuasive student writing. *Research in the Teaching of English*, 67-87.

Cooper, C. R., & Matsuhashi, A. (1983). A theory of the writing process. *The psychology of written language*, 3-39.

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

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.

Crossley, S. A., Tian, Y., & Wan, Q. (2022). Argumentation Features and Essay Quality: Exploring Relationships and Incidence Counts. *Journal of Writing Research*.

Cumming, A. (1990). Metalinguistic and ideational thinking in second language composing. *Written Communication*, *7*, 482–511.

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.

Deane, P., Quinlan, T., & Kostin, I. (2011). Automated scoring within a developmental, cognitive model of writing proficiency. *ETS Research Report Series*, *2011*(1), i-93.

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.

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.

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.

Dragsted, B., & Carl, M. (2013). Towards a classification of translation styles based on eye-tracking and keylogging data. *Journal of Writing Research*, *5*(1).

Evensen, L.S. (2002). Convention from below: Negotiating interaction and culture in argumentative writing. *Written Communication*, 19(3), 382-413.

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.

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.

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.

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.

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.

Ferris, D. R., & Hedgcock, J. (2013). *Teaching L2 composition: Purpose, process, and practice*. Routledge.

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.

Grabe, W., & Kaplan, R. B. (2014). *Theory and practice of writing: An applied linguistic perspective*. Routledge.

Graham, S., & R. Harris, K. (2000). The role of self-regulation and transcription skills in writing and writing development. *Educational psychologist*, *35*(1), 3-12.

Grömping, U. (2006). R package relaimpo: relative importance for linear regression. *J. Stat. Softw*, *17*(1), 139-147.

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.

Hadfield, J. D. (2010). MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package. *Journal of statistical software*, 33, 1-22.

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.Hammarberg, B. (2014). Problems in defining the concepts of L1, L2 and L3. *Teaching and learning in multilingual contexts: Sociolinguistic and educational perspectives, 96*, 3-18.

Harklau, L., Losey, K. M., & Siegal, M. (Eds.). (1999). *Generation 1.5 meets college composition: Issues in the teaching of writing to US-educated learners of ESL*. Routledge.

Hays, J. N., & Brandt, K. S. (1992). Socio-cognitive development and students' performance on audience-centered argumentative writing. *Constructing rhetorical education*, 202-229.

Hayes, J. R. (2012). Modeling and remodeling writing. *Written Communication*, 29(3), 369–388.

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.

Hazen, M.D. (1986). The universality of logic processes in Japanese argument. In F. Eemeren, R. Grootendorst, A. Blair, & C. Willard (Eds.), *Argumentation: Analysis and Practices* (pp. 225-237). Dordrecht-Holland: Foris Publications.

He, L., & Shi, L. (2012). Topical knowledge in ESL writing. Language Testing, 29, 443–464. Grabe, W., & Kaplan, R. B. (1996). *Theory and practice of writing: An applied linguistic perspective*. Harlow, UK: Longman.

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

Hinkel, E. (1997). Native and nonnative speakers' pragmatic interpretations of English texts. *Tesol Quarterly*, *28*(2), 353-376.

Hirose, K. (2003). Comparing L1 and L2 organizational patterns in the argumentative writing of Japanese EFL students. *Journal of Second Language Writing*, *12*(2), 181-209.

Hirvela, A. (2017). Argumentation & second language writing: Are we missing the boat? *Journal of Second Language Writing*, 36, 69–74.

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (Vol. 398). John Wiley & Sons.

Hughes, W., & Lavery, J. (2015). *Critical Thinking: An Introduction to the Basic Skills-Canadian Seventh Edition*. Broadview Press.

Hyland, K. (1996). Writing without conviction? Hedging in science research articles. *Applied linguistics*, *17*(4), 433-454.

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.

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.

Jung, J. (2017). Effects of task complexity on L2 writing processes and linguistic complexity: A keystroke logging study. *English Teaching*, *72*(4), 179-200.

Kamel, G. W. (1989). *Argumentative writing by Arab learners of English as a foreign and second language: An empirical investigation of contrastive rhetoric*. Indiana University of Pennsylvania.

Kamimura, T., & Oi, K. (1998). Argumentative strategies in American and Japanese English. *World Englishes*, *17*(3), 307-323.

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: Oxford University Press.

Knudson, R. E. (1992). Analysis of argumentative writing at two grade levels. *The Journal of Educational Research*, 85(3), 169-179.

Kobayashi, H., & Rinnert, C. (2008). Task response and text construction across L1 and L2 writing. *Journal of second language writing*, *17*(1), 7-29.

Kubota, R. (1998). An investigation of L1–L2 transfer in writing among Japanese university students: Implications for contrastive rhetoric. *Journal of Second Language Writing*, *7*(1), 69-100.

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

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).

Kunnan, A. J. (2010). Test fairness and Toulmin’s argument structure. *Language Testing*, 27, 183–189.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, 159-174.

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied linguistics*, *16*(3), 307-322.

Lee, J. J., & Deakin, L. (2016). Interactions in L1 and L2 undergraduate student writing: Interactional metadiscourse in successful and less-successful argumentative essays. *Journal of second language writing*, *33*, 21-34.

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.

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.

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).

Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: A quick and valid lexical test for advanced learners of English. *Behavior research methods*, *44*, 325-343.

Levishina, 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.

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

Lindgren, E., & Sullivan, K. P. (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.

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.

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.

Lindgren, E., & Sullivan, K. P. H. (Eds.). (2019). *Observing writing: Insights from keystroke logging and handwriting.* 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.

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.

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

McCann, T. M. (1989). Student argumentative writing knowledge and ability at three grade levels. *Research in the Teaching of English*, 62-76.

McCutchen, D. (1996). A capacity theory of writing: Working memory in composition. *Educational psychology review*, 8(3), 299-325.

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

Means, M. L., & Voss, J. F. (1996). Who reasons well? Two studies of informal reasoning among children of different grade, ability, and knowledge levels. *Cognition and instruction*, *14*(2), 139-178.

Medimorec, S., & Risko, E. F. (2016). Effects of disfluency in writing. *British Journal of Psychology*, 107(4), 625-650.

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.

Nemeth, N., & Kormos, J. (2001). Pragmatic aspects of task-performance: the case of argumentation. *Language Teaching Research*, 5, 213-240.

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.

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.

Nussbaum, E. M., & Schraw, G. (2007). Promoting argument-counterargument integration in students' writing. *The Journal of Experimental Education*, 76(1), 59-92.

Okabe, R. (1983). Cultural assumptions of East and West: Japan and the United States. In W. B. Gudykunst (Ed), *Intercultural Communication Theory: Current Perspectives* (pp. 21-44). Beverly Hills, CA: Sage Publications.

Olive, T., Alvès, R. A., & Castro, S. L. (2009). Cognitive processes in writing during pause and execution periods. *European Journal of Cognitive Psychology*, 21(5), 758-785.

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.

Olson, C. B., Scarcella, R., & Matuchniak, T. (2013). Best practices in teaching writing to English learners: Reducing constraints to facilitate writing. In S. Graham, C. A. MacArthur & J. Fitzgerald (Eds.), *Best practices in writing instruction* (2nd ed., pp. 381-403). New York: Guilford Press.

Perrin, D. (2019). Progression analysis: Working with large data corpora in field research on writing. In E. Lindgren & K. P. H. Sullivan (Eds.), *Observing Writing* (pp. 143-162). Brill.

Pessoa, S., Mitchell, T. D., & Miller, R. T. (2017). Emergent arguments: A functional approach to analyzing student challenges with the argument genre. *Journal of Second Language Writing*, *38*, 42-55.

Qin, J., & Karabacak, E. (2010). The analysis of Toulmin elements in Chinese EFL university argumentative writing. *System*, 38(3), 444-456.

Qin, J., & Liu, Y. (2021). The influence of reading texts on L2 reading-to-write argumentative writing. *Frontiers in Psychology*, *12*, 655601.

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

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.

Révész, A., Michel, M., & Lee, M. (2017). Investigating IELTS Academic Writing Task 2: Relationships between cognitive writing processes, text quality, and working memory.

Reznitskaya, A., & Anderson, R.C. (2002). The argument schema and learning to reason. In C.C. Block & M. Pressley (Eds.), *Comprehension instruction: Research-based best practices* (pp. 319-334). New York: Guilford.

Russo, J. E., Johnson, E. J., & Stephens, D. L. (1989). The validity of verbal protocols. *Memory & cognition*, *17*(6), 759-769.

Salter-Dvorak, H. (2016). Learning to argue in EAP: Evaluating a curriculum innovation from the inside. *Journal of English for Academic Purposes*, *22*, 19-31.

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.

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.

Schoonen, R., Snellings, P., Stevenson, M., & Van Gelderen, A. (2009). Towards a blueprint of the foreign language writer: The linguistic and cognitive demands of foreign language writing. *Writing in foreign language contexts: Learning, teaching, and research*, 77-101.

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.

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

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

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.Simon, S. (2008). Using Toulmin's Argument Pattern in the evaluation of argumentation in school science. *International Journal of Research & Method in Education*, 31, 277-289.

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

Spelman Miller, K. (2005). Second language writing research and pedagogy: A role for computer logging?. *Computers and Composition*, *22*(3), 297-317.

Spelman Miller, K, 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.

Stannard, R. (2019). A review of screen capture technology feedback research. *Studia Universitatis Babes-Bolyai-Philologia*, *64*(2), 61-72.

Stapleton, P. (2001). Assessing critical thinking in the writing of Japanese university students: Insights about assumptions and content familiarity. *Written communication*, 18(4), 506-548.

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.

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.

Strömqvist, 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.

Sullivan, K. P. H., & Lindgren, E. (2002). Self-assessment in autonomous computer-aided L2 writing. *ELT Journal*, 56(3), 258–266.

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

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

Tedick, D. J. (1990). ESL writing assessment: Subject-matter knowledge and its impact on performance. *English for Specific Purposes*, 9, 123–143.

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.

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). NY: 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.

Toulmin, S. E., Rieke, R. D., & Janik, A. (1984). *An introduction to reasoning*. New York: Macmillan.

Uysal, H. H. (2012). Argumentation across L1 and L2 Writing: Exploring Cultural Influences and Transfer Issues. *Vigo International Journal of Applied Linguistics*, (9), 133-159.

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.

Van Eemeren, Grootendorst, R., & Snoeck Henkemans, F. (1996). Pragma-dialectics and critical discussion. *Fundamentals of argumentation theory: A handbook of historical backgrounds and contemporary developments*, 274-311.

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.

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.

Van Waes, L., Leijten, M., Wengelin, A., & Lindgren, E. (2012). Logging tools to study digital writing processes. *Past, present, and future contributions of cognitive writing research to cognitive psychology*, 507-533.

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.

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.

Van Weijen, D., Van den Bergh, H., Rijlaarsdam, G., & Sanders, T. (2009). L1 use during L2 writing: An empirical study of a complex phenomenon. *Journal of Second language writing*, *18*(4), 235-250.

Varghese, S. A., & Abraham, S. A. (1998). Undergraduates arguing a case. *Journal of Second Language Writing*, 7, 287-306.

Walton, D. N. (1996). *Argumentation schemes for presumptive reasoning*. Psychology Press.

Wengelin, Å. (2001). Disfluencies in Writing-are they Like in Speaking?. In *ISCA Tutorial and Research Workshop (ITRW) on Disfluency in Spontaneous Speech*.

Wengelin, Å (2002) Text Production in Adults with Reading and Writing Difficulties, *Gothenburg Monographs in Linguistics 20,* Göteborg University, Department of linguistics.

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). Amsterdam, the Netherlands: Elsevier.

Wengelin, Å., 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.

Wingate, U. (2012). 'Argument' helping students understand what essay writing is about. *Journal of English for Academic Purposes*, 11, 145-154.

Xu, C. (2018). Understanding online revisions in L2 writing: A computer keystroke-log perspective. *System*, *78*, 104-114.

Yang, R. (2022). An empirical study of claims and qualifiers in ESL students' argumentative writing based on Toulmin model. *Asian-Pacific Journal of Second and Foreign Language Education*, *7*(1), 1-17.

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.

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.

Yu, V. W., & Atkinson, P. A. (1988). An investigation of the language difficulties experienced by Hong Kong secondary school students in English‐medium schools: I the problems. *Journal of Multilingual & Multicultural Development*, *9*(3), 267-284.

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.

Zhu, M., Zhang, M., & Deane, P. (2019). Analysis of keystroke sequences in writing logs. *ETS Research Report Series*, *2019*(1), 1-16.

Zhang, S. (1987). Cognitive complexity and written production in English as a second language. *Language Learning*, 37(4), 469-481.

APPENDICES

Appendix A

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

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 D

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								

Appendix E

Correlational Matric for All Numeric Variables in Linear Regression Analysis in the Entire Dataset

Variables	2	3	4	5	6	7	8	9	10	11	12	13
1. holistic scores	0.537	0.261	-0.163	-0.172	-0.244	0.093	0.307	0.293	0.377	-0.035	0.631	0.645
2. mean length of P-bursts		0.292	-0.1	-0.349	-0.518	-0.053	0.113	0.095	0.266	0.356	0.613	0.391
3. mean length of R-bursts			0.016	-0.209	-0.016	-0.463	0.141	-0.027	0.165	0.134	0.253	0.251
4. mean length of pauses				0.428	0.344	-0.084	-0.071	-0.104	-0.101	0.279	0.207	0.028
5. mean length of within-word pauses					0.311	0.175	0.052	-0.051	-0.133	-0.07	-0.069	-0.127
6. mean length of between-word pauses						-0.15	-0.164	-0.094	-0.142	-0.168	-0.175	-0.123
7. proportion of deletions							0.719	0.248	0.102	-0.229	0.134	0.096
8. mean length of deletions								0.241	0.281	-0.174	0.329	0.307
9. proportion of insertions									0.802	-0.068	0.244	0.226
10. mean length of insertions										-0.008	0.311	0.267
11. interval variance											0.155	0.042
12. typing speed												0.529
13. vocabulary knowledge												