Summer 8-12-2014

Mapping the Relationships among the Cognitive Complexity of Independent Writing Tasks, L2 Writing Quality, and Complexity, Accuracy and Fluency of L2 Writing

Weiwei Yang

MAPPING THE RELATIONSHIPS AMONG THE COGNITIVE COMPLEXITY OF INDEPENDENT WRITING TASKS, L2 WRITING QUALITY, AND COMPLEXITY, ACCURACY AND FLUENCY OF L2 WRITING

by

WEIWEI YANG

Under the Direction of Sara Weigle

ABSTRACT

Drawing upon the writing literature and the task-based language teaching literature, the study examined two cognitive complexity dimensions of L2 writing tasks: rhetorical task varying in reasoning demand and topic familiarity varying in the amount of direct knowledge of topics. Four rhetorical tasks were studied: narrative, expository, expo-argumentative, and argumentative tasks. Three topic familiarity tasks were investigated: personal-familiar, impersonal-familiar, and impersonal-less familiar tasks. Specifically, the study looked into the effects of these two cognitive complexity dimensions on L2 writing quality scores, their effects on complexity, accuracy, and fluency (CAF) of L2 production, and the predictive power of the CAF features on

L2 writing scores for each task. Three hundred and seventy five Chinese university EFL students participated in the study, and each student wrote on one of the six writing tasks used to study the cognitive complexity dimensions. The essays were rated by trained raters using a holistic scale. Thirteen CAF measures were used, and the measures were all automated through computer tools. One-way ANOVA tests revealed that neither rhetorical task nor topic familiarity had an effect on the L2 writing scores. One-way MANOVA tests showed that neither rhetorical task nor topic familiarity had an effect on accuracy and fluency of the L2 writing, but that the argumentative essays were significantly more complex in global syntactic complexity features than the essays on the other rhetorical tasks, and the essays on the less familiar topic were significantly less complex in lexical features than the essays on the more familiar topics. All-possible subsets regression analyses revealed that the CAF features explained approximately half of the variance in the writing scores across the tasks and that writing fluency was the most important CAF predictor for five tasks. Lexical sophistication was however the most important CAF predictor for the argumentative task. The regression analyses further showed that the best regression models for the narrative task were distinct from the ones for the expository and argumentative types of tasks, and the best models for the personal-familiar task were distinct from the ones for the impersonal tasks.

INDEX WORDS: Cognitive complexity, Cognitive demand, Second language writing, Writing performance, Writing assessment, Complexity, accuracy, and fluency, Task-based language teaching

MAPPING THE RELATIONSHIPS AMONG THE COGNITIVE COMPLEXITY OF INDEPENDENT WRITING TASKS, L2 WRITING QUALITY, AND COMPLEXITY, ACCURACY AND FLUENCY OF L2 WRITING

by

WEIWEI YANG

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

2014

MAPPING THE RELATIONSHIPS AMONG THE COGNITIVE COMPLEXITY OF INDEPENDENT WRITING TASKS, L2 WRITING QUALITY, AND COMPLEXITY, ACCURACY AND FLUENCY OF L2 WRITING

by

WEIWEI YANG

Committee Chair: Sara Weigle

Committee: Diane Belcher
Eric Friginal
YouJin Kim
T. Chris Oshima

Electronic Version Approved:

Office of Graduate Studies
College of Arts and Sciences
Georgia State University
August 2014

DEDICATION

To my family

ACKNOWLEDGEMENTS

The completion of this dissertation would not have been possible without the support of many individuals, the many learning opportunities in my academic department and beyond, and two professional organizations. I would like to particularly thank my advisor, Dr. Sara Weigle, for her constant support and encouragement throughout my dissertation stage and my entire PhD study period. Her feedback on the dissertation project and the manuscript is highly valuable and constructive. The level of detailedness and thoroughness of her comments have encouraged me to perfect my work to the extent possible. I also greatly appreciate her willingness to let me use her Educational Testing Service (ETS) essay data for several course projects I worked on, and these projects have doubtlessly boosted my ability and confidence in working on the types of linguistic analyses and data analyses required for the dissertation work. The several opportunities to work with her on research projects and publications have also sharpened my skills as a researcher and writer.

I would also like to extend my sincerest gratitude to each of the members of my dissertation committee. First, I would like to thank them all for their constructive comments on my dissertation project and the manuscript, which with no doubt strengthened the quality of the research work and the reporting. Additionally, I have greatly benefited from the quantitative research methods courses offered by Dr. Chris Oshima and Dr. Eric Friginal, without which I would not have been able to confidently do the statistical analyses for this project. Many thanks to Dr. Oshima for patiently answering my endless quantitative data analysis questions for my research projects. Special thanks to Dr. Friginal for providing constructive comments on a research project related to this dissertation work.

I feel grateful to Dr. Diane Belcher for sharing with me her insights into second language writing that made me think deeper into the writing contexts and issues examined in my dissertation. I am also thankful for her support and encouragement for my publication of a forum article in *TESOL Quarterly*, which helped me to further develop knowledge and expertise on an important construct examined in this project. Finally, many thanks to Dr. YouJin Kim for sharing with me her insights into task-based language teaching in relation to my dissertation, and I feel very fortunate to have her on my committee for her expertise on and genuine interest in the puzzling and fascinating questions my work examines. I also benefited from working with her on several of her research projects and further honed my own research skills.

The dissertation work would not have been completed without the support of many of my colleagues and friends. In particular, I would like to thank Susan Firestone, Sarah Goodwin, Merideth Hoagland, Pamela Pearson, Nic Subtirelu, and HaeSung Yang in my academic department for their willingness to help with rating the essays, even though I knew how busy they were with their own work. I would also like to thank Dr. Lei Shen of Nanjing University for her help with scoring of the cloze tests and my friend Lin Dong for her help with initial data management and some linguistic analyses. I am indebted to Dr. Xiaofei Lu at Penn State University for his willingness to process my essay data with his computational tools for most of the lexical and syntactic complexity analyses of this study and for his patience in answering many of my questions about the tools. Finally, I am extremely obliged to my friend Hongcheng Dong and Dr. Quanwang Zhan at the participating university in China for their help with contacting and coordinating with the English teachers in the university for the recruitment of the participants. Sincere thanks also go to the English teachers at the participating university who administered the research materials and the students who willingly participated in the study.

Finally, I would like to thank The International Research Foundation (TIRF) for funding the research and my travels to two conferences to present the study through its Doctoral Dissertation Grant. I would also like to thank Educational Testing Service (ETS) for processing my essay data for the accuracy indices with the e-rater engine and particularly Chaitanya Ramineni at the Assessment Innovations Research Center of ETS for patiently dealing with my request and meeting with my specific needs for the data.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

CAF: complexity, accuracy, and fluency

TBLT: task-based language teaching

RT: rhetorical task

SC: syntactic complexity

CHAPTER 1: INTRODUCTION

1.1 Context of the Study

In language development and language assessment, the cognitive complexity of language tasks can play an important role in the quality of task performance and in the language features produced for fulfilling the tasks. An understanding of key cognitive complexity factors and the roles they play in learner performance can enable language teachers and materials writers to design and sequence language tasks more properly so that learners progress along tasks of increasing complexity and develop their interlanguage progressively. Such an understanding can also provide valuable information to language assessment specialists about proper tasks to use for assessing task performance of specific learner populations so that learners can demonstrate their ability at their highest capacity and their interlanguage can be properly assessed. To these practical ends, fortunately, language researchers from main two areas have pursued work related to the cognitive complexity of tasks and its effect on learner performance: first and second language writing and task-based language teaching (TBLT).

In the literature, the cognitive complexity of tasks, otherwise termed the cognitive demands of tasks, is defined as "the level of thinking skills or intellectual functioning required to accomplish certain tasks" (Hale et al., 1996, p. 12), “the extent to which task characteristics can affect the allocation of an individual’s attention, memory, reasoning and other processing resources” (Robinson, 2007a, p. 17), or more simply “the thinking required … for a task” (Skehan, 1998, p. 99). Although there is a body of work examining the relationship between the cognitive complexity of tasks and task performance in both the writing literature and the TBLT and some of the research questions pursued in the two areas overlap, it appears that researchers in the two areas rarely cite each other’s work, even though certain cognitive complexity factors are examined in both areas. The current study investigates the effects of the cognitive complexity of independent writing tasks on college-level L2 writing performance in terms of writing quality and features of writing in the areas of linguistic complexity, linguistic accuracy, and fluency (collectively known as CAF), as well as the predictive power of the CAF features on writing quality for tasks of different cognitive complexity, by drawing on and bridging the research and findings from the two areas. The cognitive complexity dimensions studied in the current study are rhetorical task and topic familiarity.

Task-based language teaching (TBLT) is a second language (L2) teaching approach that developed with increased popularity of communicative approaches to L2 teaching and learning from the 1980s, where communication is seen as the primary means and end of language development. In TBLT, real-world tasks involving use of language and pedagogical tasks that are seen to promote learners' ability to perform in real-world tasks are used as the basic units to organize the syllabus, the instruction and the assessment (Ellis, 2003; Long & Crookes, 1993; Skehan, 1998), instead of using linguistic components such as grammatical rules or using language functions and notions for such purposes. In the TBLT literature, there is explicit theorizing of the effects of the cognitive complexity of tasks on language performance in the CAF areas, and there are competing theories about such effects, which will be explained in relation to the relevant cognitive complexity dimensions examined in this study briefly in this chapter and in detail in Chapter 2. The TBLT framework of teaching and assessment applies to both spoken and written tasks, and the theorizing of the effects of task cognitive complexity on task performance has not been formally distinguished for performance in spoken and written modalities.

The first cognitive complexity dimension researched in this study is rhetorical task. Rhetorical task, with narrative, expository, and argumentative tasks examined the most often, is the most studied cognitive complexity dimension in the first and second language writing literature. Rhetoric theories (e.g. Bain, 1967; Brooks & Warren, 1979; Cairns, 1899; Genung, 1900; Moffett, 1968), taxonomies of educational objectives (e.g., Bloom, 1956; Anderson & Krathwohl, 2001), and human cognitive development trajectories (e.g., Kuhn & Franklin, 2006; Piaget, 1972) all suggest that the three more commonly examined rhetorical tasks pose different levels of cognitive demands on writers, with personal narration the least cognitively demanding, exposition cognitively more demanding than narration, and argumentation the most cognitively complex. In general, these tasks differ in terms of types of thinking involved and thus inherently different levels of cognitive demands (Hale et al., 1996; Moffett, 1968; Weigle, 2002), as well as whether reasoning is required and the degree of reasoning called for (Bain, 1967; Brooks & Warren, 1979; Cairns, 1899; Genung, 1900). Empirical studies of first language (L1) writing have shown that rhetorical task significantly affects writing quality for certain age groups (e.g., Kegley, 1986; Quellmalz, Capell, & Chou, 1982) and language features of writing (e.g., Crowhurst, 1980; Ravid, 2004).

Although rhetorical task is the most studied cognitive complexity dimension in writing studies and has been proven to affect task performance, it cannot find a clear representation in the two main cognitive complexity frameworks in the TBLT literature – Robinson (2007a) and Skehan (1998). But with their inherently different levels of cognitive demands and their different levels of involvement of reasoning in the different task types, rhetorical task seems to rather fit into the resource-directing category in Robinson’s framework for task complexity (cognitive factors) where there are factors regarding whether and the extent to which reasoning (e.g., causal reasoning and intentional reasoning) and perspective-taking are involved. When rhetorical task is analyzed based on Robinson's framework and Cognition Hypothesis (Robinson, 2001; 2003; 2005; 2007a; 2010), it is predicted that learners will produce language of higher accuracy and linguistic complexity but lower fluency when performing on the more complex tasks, since learners' attentional and memory resources will be directed to form-function mappings due to the inherent cognitive/conceptual demands in complex tasks. However, Skehan's Trade-off Hypothesis (Skehan, 1992; 1996; 1998; Skehan & Foster, 2001) would not make such predictions for the beneficial effects of complex rhetorical tasks on accuracy and complexity, but rather predict lower accuracy, complexity, and fluency in performance on complex tasks (Robinson & Gilabert, 2007). Certainly, L2 studies examining rhetorical task through these lens are much needed to test the competing hypotheses in the TBLT literature. With that stated, L1 writing studies have already suggested higher syntactic complexity (e.g., Beers & Nagy, 2007; Crowhurst, 1980), higher lexical complexity (Ravid, 2004), but lower fluency (Beers & Nagy, 2007; San Jose, 1972) and lower accuracy (Pringle & Freedman, 1979) in expository and argumentative essays than those in narrative essays, partially supporting Robinson's hypothesis. How these effects are played out in L2 task performance requires more empirical investigations.

Different levels of cognitive demands intrinsic in narrative, expository, and argumentative tasks affect not only language production, but also quality of writing in terms of scores granted – certainly a construct highly related to the language produced in writing. The effect of rhetorical task on writing quality, though, seems to be more dependent on writers' age groups and amount of writing experience. L1 writing studies have suggested the trajectory of better performance in narrative tasks by younger L1 writers (e.g., Freedman & Pringle, 1984; Kegley, 1986; Prater & Padia, 1983) and then equally good or even better performance in expository and argumentative tasks from high school or so (e.g., Prater, 1985; Quellmalz, Capell, & Chou, 1982). Fewer L2 writing studies have examined the relationship, mostly examining college-level writing only, and the findings from the few are rather inconclusive (Carlson, Bridgeman, & Waanders, 1985; Hamp-Lyons & Mathias, 1994; Lim, 2009; Park, 1988; Spaan, 1993), possibly due to different topics used and/or unclear task classifications. Adult L2 writers are probably cognitively mature enough to handle all the rhetorical tasks, given that they have adequate prior writing experience with the tasks. But it still remains to be tested whether they would perform differently across the different rhetorical tasks in terms of scores and to what extent each rhetorical task can distinguish higher- and lower- performers.

L1 and L2 writing studies have painted the picture that rhetorical task effects writing performance, but the picture is not near complete without a consideration of topics used for the writing. For any rhetorical task, different topics can be used. Probably just as much as rhetorical task, different topics mean difference in task performance even for topics of the same rhetorical task, in terms of scores (e.g., Calman, 1986; Gabrielson, Gordon, & Engelhard, 1995; Hamp-Lyons & Mathias, 1994; Tedick, 1990) and language production features (e.g., Nold & Freedman, 1977; Tedick, 1990; Yang, Lu, & Weigle, 2012; Yang & Weigle, 2011). How topics can be grouped based on their cognitive complexity for writers is another dimension that must be taken into account when we consider the cognitive complexity of writing tasks. Topic familiarity is the only cognitive complexity dimension that is explicitly shared in the writing literature and the two main cognitive complexity frameworks in the TBLT literature (Robinson, 2007a; Skehan, 1998). Further, Skehan (1998) aligns personal vs. impersonal tasks with the topic familiarity dimension, with personal seen as more familiar and impersonal less familiar. However, although writers typically have the greatest knowledge and familiarity with personal topics in common life domains, impersonal topics can vary in terms of the amount of knowledge and familiarity writers have on the topics as well, with some more familiar and others less familiar. Topic familiarity can be categorized, in descending knowledge or familiarity order, into personal-familiar, impersonal-familiar, and impersonal-less familiar topics. Based on dual processing theories (Evans, 2010; Evans, 2011; Stanovich, West, & Toplak, 2011), highly compiled knowledge built through rich experience allows more autonomous and less effortful cognitive processing, thus making task performance cognitively less demanding, in comparison to low knowledge or familiarity about content and task, which requires more reflective and effortful cognitive processing and leads to higher cognitive load.

On this cognitive complexity dimension of topic familiarity, Robinson's Cognition Hypothesis (Robinson, 2001; 2003; 2005; 2007a; 2010) and Skehan's Trade-off Hypothesis (Skehan, 1992; 1996; 1998; Skehan & Foster, 2001) have the same predictions for its effects on language production in the CAF areas, with higher familiarity leading to higher accuracy, complexity, and fluency, although Skehan predicts some types of trade-off among the three performance areas, and lower familiarity resulting in lower accuracy, complexity, and fluency. Although no empirical study has compared all the three levels of topic familiarity (i.e., personal, impersonal-familiar, and impersonal-less familiar topics), in several L2 writing studies examining mostly college-level ESL writing, personal topics, compared to impersonal topics in a general sense, have been found to elicit linguistically less complex language production (Hinkel, 2002; Spaan, 1993; Yu, 2010) but longer texts and higher linguistic accuracy (Spaan, 1993), and impersonal familiar topics, in comparison to impersonal less familiar topics, are yet found to invite lexically and syntactically more complex language production (Tedick, 1990; Yu, 2010) and longer texts (Tedick, 1990). Taken together, these L2 studies suggest the inadequacy of a simple dichotomy of personal vs. impersonal topics and the need to analyze impersonal topics on the familiarity dimension as well. Further, regarding the effects of these three levels of familiarity on the CAF performance areas, although not much can be said about the effects on accuracy and fluency based on the above L2 studies, it appears that the highest linguistic complexity is achieved at the medium familiarity level–the impersonal familiar one, a conclusion that challenges the predictions drawn based on Robinson's and Skehan's hypotheses that the highest complexity is found for personal topics which mean the greatest knowledge and familiarity to writers. The extent to which the predictions made in the TBLT literature are correct can probably be best tested in empirical studies that examine the topic familiarity dimension for all the three levels–personal, impersonal-familiar, and impersonal-less familiar topics.

The story for the effect the topic familiarity on writing scores is quite similar. Studies examining adult ESL writing suggest higher scores for impersonal topics in a general sense than those for personal topics (Hamp-Lyons & Mathias, 1994; Hinkel, 2002; Yu, 2007) and higher scores for impersonal familiar topics than those for impersonal less familiar topics (Tedick, 1990), thus suggesting the highest writing quality produced by this population for medium familiarity topics, the impersonal familiar ones. However, it is unknown whether there would be any performance difference for personal topics and impersonal less familiar topics. Studies comparing score differences for all the three levels of topic familiarity are certainly welcome.

As outlined above, researchers have invested efforts in understanding the role of cognitive complexity in task performance in terms of language production features and the quality of task performance; however, the connections between language production features and performance quality as operationalized by scores for tasks of different cognitive complexity have not yet been established. In the context of L2 writing, the questions are to what extent language production features predict writing quality for tasks posing different cognitive loads to writers and whether the language production features have different predictive values on writing quality when the cognitive complexity of tasks varies along a certain dimension. For instance, expository and argumentative tasks have been found to elicit syntactically and lexically more complex language than narrative tasks (e.g., Beers & Nagy, 2007; Crowhurst, 1980; Lu, 2011; Ravid, 2004); does this mean that syntactic and lexical complexity will be able to predict more of the writing quality for expository and argumentative tasks than that for narrative tasks? All these questions beg more empirical investigations. Answers to these questions will have implications for the development of essay scoring rubrics and the design of automated essay scoring systems where language production features of essays are often an important part. Answers to these questions will also have implications for teaching of L2 writing which can benefit from an understanding of key language production features that are important to writing quality for different task types.

1.2 Research Questions

The current study pursues the following six research questions, with the first three questions addressing rhetorical task and the last three addressing topic familiarity.

1. What is the effect of varying the cognitive complexity of independent writing tasks along the rhetorical task dimension on L2 writing scores of college-level ESL writers'?
2. What are the effects of varying the cognitive complexity of independent writing tasks along the rhetorical task dimension on linguistic complexity, accuracy, and fluency of L2 writing production of college-level ESL writers'?

3. How do linguistic complexity, accuracy, and fluency of L2 writing production predict L2 writing scores of college-level ESL writers' for independent writing tasks of different cognitive complexity along the rhetorical task dimension?
4. What is the effect of varying the cognitive complexity of independent writing tasks along the topic familiarity dimension on L2 writing scores of college-level ESL writers'?
5. What are the effects of varying the cognitive complexity of independent writing tasks along the topic familiarity dimension on linguistic complexity, accuracy, and fluency of L2 writing production of college-level ESL writers'?
6. How do linguistic complexity, accuracy, and fluency of L2 writing production predict L2 writing scores of college-level ESL writers' for independent writing tasks of different cognitive complexity along the topic familiarity dimension?

1.3 Operationalizations of Rhetorical Task and Topic Familiarity in the Study

In the current study, rhetorical task is operationalized into four levels: narration, exposition, expo-argumentation, and argumentation. Expo-argumentation, a category in between the traditionally used categories of exposition and argumentation, was created and studied on two grounds. First, this middle-ground category corresponds with Genung's (1900) category of "exposition of the symbols of things" which is about generalizations of what things mean and involves interpretation and criticism, a type of exposition Genung distinguishes from "exposition of things" which is about generalizations of actual things. Second, in actual writing studies, the classifications for exposition and argumentation have not been highly consistent, particularly with writing involving personal judgment and opinions but not involving taking a side on something debatable sometimes classified as exposition and other times as argumentation (e.g., Hamp-Lyons & Mathias, 1994; Spaan, 1993). Exposition and argumentation are also sometimes used to denote one rhetorical task and to mean the same thing (See Crowhurst, 1990; Ravid, 2004). The middle-ground category of expo-argumentation is used to address the inconsistencies observed. The following operational definitions of exposition, expo-argumentation, and argumentation are adopted for the current study and are used to examine and interpret previous related work in the literature that is reviewed in the next chapter:

Exposition: A rhetorical task that mainly invites a writer to explain and to provide information about something (not to give personal opinions or judgment or to argue on the topic), based on facts and generalizations of events and states.

Expo-argumentation: A rhetorical task that mainly invites a writer to explain, to provide information about something, with personal opinions and judgment on the topic involved (but not to take a stand on a debatable issue or statement), based on facts, generalizations, and reasoning.

Argumentation: A rhetorical task that mainly invites a writer to give personal opinions and judgment on a debatable issue or statement and to take a stand on the issue/statement, based on facts, generalizations, and reasoning.

These three levels of rhetorical task– exposition, expo-argumentation, and argumentation, with their operational definitions and sample prompts, were validated through a pilot study which is reported in detail in Appendix A.

Topic familiarity is defined in the current study as the amount of direct and explicit knowledge a writer presumably has in relation to a topic, with knowledge built from different kinds of experience such as personally physically experiencing or observing something, conversing and thinking about something, and obtaining information about something from other people or knowledge sources. Topic familiarity is operationalized into three levels in this study:

higher familiarity (i.e., personal familiar topics), medium familiarity (i.e., impersonal familiar topics), and lower familiarity (i.e., impersonal less familiar topics). A personal topic means one for which writers are invited to write about themselves and the relationship of the writing subject matter with their own life. An impersonal topic refers to one for which writers are invited to write about a group or people in general in relation to a subject matter or about objects and abstract concepts.

Writing subject matters, topics, and prompts are distinguished as follows. A subject matter refers to the subject/ thing being written on, such as traditional food, dance, and exercise. A topic means what is exactly construed in a writing prompt, i.e., what is specifically asked or stated about a subject matter, such as how to cook a Chinese traditional food, how dance represents a country's culture, or what role exercise can play in helping children with obesity. And a writing prompt covers not only the subject matter and the topic written on but also the actual wording used, including word choice, syntax, number of words, and so on.

CHAPTER 2: REVIEW OF RELATED LITERATURE

In this chapter, related literature about the cognitive complexity dimensions of rhetorical task and topic familiarity and their effects on writing performance will be reviewed in greater details, as well as theories of cognitive complexity and its effects on complexity, accuracy, and fluency (CAF) of language production in the TBLT literature and related studies on the cognitive complexity dimensions under focus in that line of literature. In addition, measures for CAF used in TBLT studies and L2 writing studies are summarized and discussed in view of the constructs being measured.

2.1 Rhetorical Task, Cognitive Complexity, and Writing Performance

2.1.1 Rhetorical task and cognitive complexity

Rhetorical task (RT) is also known in L1 and L2 writing development as rhetorical mode and discourse mode (Moffett, 1968; Weigle, 2002) and in rhetoric as form of composition (Cairns, 1899), literary type (Genung, 1900), form of discourse (Corbett, 1965), kind of composition (Bain, 1967), kind of discourse (Brooks & Warren, 1979). It typically includes four task types: narration, description, exposition, and argumentation. These distinctions of rhetorical task have often been based on their different rhetorical purposes and functions (Bain, 1967; Brooks & Warren, 1979; Cairns, 1899; Genung, 1900; Kinneavy, 1971; Smith, 2003). Genung (1900) specifies that narration regards "invention dealing with events", description "invention dealing with observed objects", exposition "invention dealing with generalized ideas", and argumentation “invention dealing with truths, and with issues of conviction" (p. 475). Genung further divides exposition into two types: 1) exposition of things which is about generalizations of actual things, and 2) exposition of the symbols of things which is about generalizations of what things mean and involves interpretation and criticism. Defining and expounding the nature of an actual thing based on generalized facts and principles belongs to the first category of exposition in Genung. The literary work of the essay which involves personal interpretation and opinion belongs to the second category of exposition in Genung and is “merely to open questions, to indicate points, to suggest cases, to sketch outlines” (Morley, not dated, cited in Genung, 1990, p. 595). Although there are several rhetorical tasks, classical rhetoric is largely concerned with the art and strategies of persuasion into a belief or an action (Corbett, 1965; Smith, 2003).

In the current study, the rhetorical tasks of narration, exposition, and argumentation are all examined. Similar to the distinction that Genung (1900) makes, the current study has the exposition category as exposition of actual things and creates a category named expo-argumentation to mean the exposition of the symbols of things in Genung. The term expo-argumentation is used partly because of the more broadened view of what argumentation includes in modern rhetoric (e.g., Lunsford, Ruszkiewicz, & Walters, 2001) and partly because of the use of argumentation to mean open-ended opinion essays not asking for a stand on something debatable in some L2 writing studies (e.g., Hamp-Lyons & Mathias, 1994; Spaan, 1993). Thereby, the four rhetorical tasks examined in the current study are named narration, exposition, expo-argumentation, and argumentation, with argumentation dealing with debatable issues in which a stand is required. Working definitions for exposition, expo-argumentation, and argumentation can be found in the final section of Chapter 1.

Coming to the cognitive complexity these four rhetorical tasks entail, there are different types of thinking involved in performance of the different task types, creating different amount of cognitive demands, with narration, exposition, expo-argumentation, and argumentation forming a cline of increasing cognitive complexity for writers. Rhetoric theories (e.g. Bain, 1967; Brooks & Warren, 1979; Cairns, 1899; Genung, 1900; Moffett, 1968), taxonomies of educational objectives (e.g., Bloom, 1956; Anderson & Krathwohl, 2001), and human cognitive development trajectories (e.g., Kuhn & Franklin, 2006; Piaget, 1972) all lend support for the different levels of cognitive demands inherent in the rhetorical tasks. In general, personal narrative tasks require recalling and retrieving events (e.g., Cairns, 1899; Moffett, 1968), exposition requires recalling and generalizing based on events and states (e.g., Bain, 1967; Smith, 2003), and expo-argumentation and argumentation involve recalling, generalizing, and reasoning, with argumentation demanding more reasoning (e.g. Bain, 1967; Brooks & Warren, 1979; Cairns, 1899; Genung, 1900). Generalizing is limited here to deriving truths and general principles from events and states. Reasoning, a general term to refer to the cognitive processes involved in making personal interpretations and judgments about something using reasons and logic, often includes the two broad categories of inductive reasoning and deductive reasoning (e.g., Genung, 1900, Cairns, 1899). The reasoning demanded in expo-argumentation and argumentation is also called inferencing (Genung, 1900; Brooks & Warren, 1979); Genung (1900) also uses the terms of "inferencing from particulars" and "inferencing from generals" to mean inductive and deductive types of reasoning. Argumentation is seen to demand greater reasoning in comparison to expo-argumentation, not only because to establish an argument, various reasoning processes are often employed, but also because to make a defendable and strong argument, alternative views often need to be addressed and tackled in some ways through reasoning (Genung, 1900). As Brooks and Warren (1979) states, reasoning/ inferencing is "fundamental to the intention of the discourse [of argumentation]" (p. 131)

The different cognitive demands intrinsic in the different rhetorical tasks are also aligned with classifications of cognitive processes for educational objectives (e.g., Bloom, 1956), in which cognitive processes form different levels of cognitive demands and are used as the basis for setting up learning procedures. For example, in Bloom's (1956) original taxonomy of educational objectives and its revision by Anderson and Krathwohl (2001), recalling information from long-term memory is a cognitive process at the lowest level in terms of cognitive demands, and analysis and evaluation are cognitive processes at higher levels regarding cognitive demands, with evaluation at a higher level than analysis. Personal narrative tasks primarily rely on the cognitive process of recalling events. Exposition mainly requires the cognitive processes of recalling information and analyzing the information recalled, and it is through analysis that we derive generalizations (Bain, 1967; Genung, 1900). For expo-argumentation and argumentation, not only information retrieval and analysis but also evaluation of information and ideas is needed. According to Anderson and Krathwohl, evaluation is the process of "[making] judgments based on criteria and standards" (2001, p. 68), which exactly depicts the essential cognitive functioning required for expo-argumentative and argumentative tasks. In the writing literature, cognitive demands of rhetorical tasks have been associated with the cognitive processes presenting different levels of cognitive demands specified in Bloom and others (e.g., Hale et al., 1996; Weigle, 2002).

2.1.2 Rhetorical task and first and second language writing performance

2.1.2.1 Rhetorical task and writing scores

Regarding any difference in writing performance in terms of scores for the different RTs, L1 studies have mostly compared performance in narrative and argumentative essays, with fewer examining performance difference between narration and exposition, and a handful of L1 and L2 studies have examined difference in performance in expository or expo-argumentative and argumentative essays.

2.1.2.1.1 Narration vs. argumentation or exposition

A number of L1 studies compared writers' performance in narration and argumentation, with fewer examining performance difference between narration and exposition First, there is rather strong and converging evidence from the L1 studies that L1 writers receive higher scores on narrative tasks than those on argumentative tasks across grade levels, from elementary school (e.g., Prater & Padia, 1983; Sachse, 1984), to middle school (e.g., Kegley, 1986; Freedman & Pringle, 1984), and into high school years (e.g., Calman, 1986; Prater, 1985). Although no study could be identified for a comparison in scores in narration and argumentation for college L1 writers, Raimes' (1987) study of composing processes in the two modes suggests that, even for university-level students, argumentation is cognitively more demanding than narration. Comparing performance on narrative and expository tasks, the fewer studies examining them reported lower scores in exposition than narration for L1 writers in elementary school (Prater & Padia, 1983) and middle school (Engelhard et al., 1992; Kegley, 1986), but for high school L1 writers, both lower performance (Prater, 1985) and higher performance (Quellmalz, Capell, & Chou, 1982) have been reported for expository tasks in comparison to narrative tasks. In general, the performance differences for L1 writers in the different RTs seem to be due to the higher cognitive demands of argumentative and expository tasks, as well as the kinds of writing practice and instruction the students receive in school (Engelhard et al., 1992; Prater & Padia, 1983) and potentially raters' differential judgment towards the tasks (Quellmalz, Capell, & Chou, 1982). It should also be noted that these L1 studies were mostly conducted in the 1980s, and the status quo of students' writing performance in the different RTs in more recent years is not known from the literature.

For ESL writers' writing quality in the modes of narration and argumentation or exposition, only one relevant study could be located. Lim (2009) examined 2003-2008 MELAB (Michigan English Language Assessment Battery) essays written mostly by adult ESL writers and grouped the 57 prompts used into 5 narration, 30 exposition, and 22 argumentation. No definitions for the RTs or actual prompts for the RTs are shown in the report, though. Lim found no statistical difference in writing quality across the RTs, but narration received lower scores in comparison to exposition and argumentation. With only this ESL study which primarily examined adult ESL writing, we certainly cannot draw any conclusion for ESL writers' performance difference in narration and argumentation or exposition. For future inquiries, ESL writers' age groups and prior writing practice should be considered for such comparisons.

2.1.2.1.2 Exposition or expo-argumentation vs. argumentation

In terms of a comparison of writing quality for expository or expo-argumentative and argumentative essays, several L1 and L2 studies are relevant here. First, based on the several L1 studies, L1 writers produce writing of higher quality for exposition than argumentation, in elementary school (Prater & Padia, 1983) and middle school (Kegley, 1986), but their writing quality in the two rhetorical tasks was not found to differ in high school (Prater, 1985). The findings of these L1 studies again support the much repeated observation that argumentation is the most difficult in comparison to narration and exposition.

In L2 studies, Carlson, Bridgeman, and Waanders (1985) examined performance differences in expository and argumentative writing by adult ESL writers with intermediate to intermediate-high L2 proficiency. This study and Park (1988)–another study using part of the data from Carlson, et al.–reveal that topic is a factor that affects the performance difference in the two RTs for adult ESL writers. Carlson, et al. collected essays of the two RTs by having 662 college-level EFL writers and 55 college-level L1 writers each write on four topics, with two expository essays based on information from graphs (U.S. farming and seven continents) and two argumentative essays (exploration of outer space and physical vs. intellectual ways of leisure). The authors report a moderate to high correlation between holistic scores for the two different RTs for the whole sample (r =.83, $r_{adjusted}$ = 1.0); however, no statistical testing was reported for the mean differences. The descriptive statistics reported for the ESL writers, however, show that their performance in the two RTs varied as a function of topic. The writers performed better in the leisure argumentative task than the two expository tasks but performed worse in the outer space argumentative task than the two expository tasks. Park (1988), using some of writing data from Carlson, et al. (1985) for the outer space argumentative task and the U.S. farming expository task, reported significantly higher holistic scores for the expository task than the argumentative one.

As for a comparison in performance on expo-argumentative and argumentative types of tasks, two L1 studies involving college-level writers are pertinent, and they both suggest that there may be no performance difference for the two rhetorical tasks. Greenberg (1981), examining 192 college freshmen’s expo-argumentative writing which asked for interpretations on certain issues and argumentative writing which asked for a position on the same issues by agreeing or disagreeing with provided statements, found that the students performed equally well on the two RTs. In a composing processes study, Witte (1987) is able to provide some evidence for why expo-argumentation appears to be as demanding as argumentation, even though the reasoning demands in argumentation are generally higher. Through an examination of composing processes in the two RTs of college L1 writers who were mostly freshmen, it is revealed that these two RTs present their own particular challenges to writers. Witte notes, “Compared to the argumentative tasks, the expository writing tasks are very open-ended with respect to rhetorical situation, with respect to suggested discourse schema, and with respect to usable ideational content." (p. 420), and acknowledges the challenges posed to writers due to the more open-ended nature of expo-argumentative tasks.

In L2 studies, three studies that all used the MELAB essay-writing data examined expo-argumentative types of writing in comparison to argumentative writing; however, in these studies, it seems that some expo-argumentative prompts are classified as expo-argumentation which the authors call exposition and others as argumentation. For example, in Spaan (1993), the prompt question of “Pick one [energy] source and discuss the situation for which it is best suited” is classified as argumentation. And in both Spaan (1993) and Hamp-Lyons and Mathias (1994), the open-ended prompt question of “What is your opinion of mercenary soldiers …? Discuss” is classified as argumentation. Further, in Hamp-Lyons and Mathias, narrative and descriptive tasks are all grouped under the expo-argumentative category which the authors call expository. As for the findings of the studies, Spaan (1993) and Lim (2009) did not find a statistical difference in the writing scores for the expo-argumentative and argumentative tasks, although scores for expo-argumentation were higher than those for argumentation in both cases, whereas Hamp-Lyons and Mathias reported significantly higher scores for argumentation than expo-argumentation. The conflicting findings reported in the studies may be due to the somewhat unclear classifications of expo-argumentative and argumentative tasks and the inclusion of narrative and descriptive tasks in Hamp-Lyons and Mathias. It is plausible that, as the college-level L1 writing studies indicate, there is no performance difference in the two RTs in terms of scores for adult ESL writers with intermediate to intermediate-high L2 proficiency. For younger

L1 and L2 writers, as well as the most mature and advanced ones, there is a lack of empirical studies studying their performance in the two RTs.

In sum, the few L1 and L2 studies suggest that performance difference in expository, expo-argumentative, and argumentative tasks differs across school grades and language proficiency levels. L1 writers at K-12 levels tend to do better in expository tasks in comparison to argumentative tasks (Prater & Padia, 1983; Kegley, 1986), but they may fare equally well in the two RTs in high school (Prater, 1985). Although no study compared college-level L1 writing in exposition and argumentation, the two L2 studies with adult ESL writers (Carlson, et al., 1985; Park, 1988) indicate that adult writers can perform well on both of the RTs, but topic is a factor that may affect their performance. Further, writers' performance in expo-argumentative and argumentative tasks, the two most demanding RTs, may not differ, at least during early college years (Greenberg, 1981; Lim, 2009; Spaan, 1993; Witte, 1987).

Overall, studies looking into performance difference in different rhetorical tasks suggest writers' trajectory of developing competence along the RT complexity spectrum: narration, exposition, and argumentation, as writers mature in age and develop in language and writing. For the adult college-level ESL writers the current study focuses on, it can be hypothesized that they will perform equally well on the four rhetorical tasks.

2.1.2.2 Rhetorical tasks and essay language production features

Studies of learners' development in writing on different rhetorical tasks have also examined the relationship between RTs and essay language production features, notably syntactic complexity. Few studies have looked into other language production features such as total number of words generated, lexical sophistication, and accuracy of language production. Measures of syntactic complexity (SC) are largely based on Hunt's (1965) work examining syntactic maturity in writing of L1 writers across grade levels and age groups. The most commonly used SC measures are mean length of T-unit (MLTU), clauses per T-unit (C/TU), and mean length of clause (MLC). T-unit is proposed by Hunt as "minimal terminable unit" (p. 21) and is defined as “one main clause with all the subordinate clauses attached to it” (p. 20). C/TU is proposed by Hunt to refer to a measure for the amount of subordination, but clause in Hunt’s work and the L1 and L2 writing studies in the writing literature only includes finite clause as seen in the definition Hunt provides for clause - “a structure with a subject and a finite verb (a verb with a tense marker)” (p.15). Thereby, C/TU in the writing literature essentially means the amount of finite subordination, excluding in the picture nonfinite subordination such as infinitive and gerund used as complement and infinitive and participle used as adverbial for sentences or modifier for nouns. And MLC in the writing studies actually refers to the length of finite clauses, with nonfinite elements counted in such a measure.

2.1.2.2.1 Narration vs. argumentation, exposition or expo-argumentation

L1 writing studies point to a rather converging conclusion that argumentative and expository discourse is syntactically more complex than narrative discourse. Several L1 studies compared SC of L1 essays in the modes of narration and argumentation, and they all reported significantly higher SC in argumentative writing than that in narrative writing. The finding is borne out in San Jose’s (1972) study with fourth graders in MLTU and C/TU but not MLC, Beers and Nagy’s (2007) study with seventh and eighth graders in MLTU and C/TU but not MLC, Crowhurst’s (1980a) study with sixth, tenth, and twelfth graders in MLTU for each grade, and Crowhurst and Piche’s (1979) study with sixth and tenth graders in MLTU and C/TU for each grade and in MLC for tenth grade but not sixth grade. San Jose also compared SC in expository and narrative writing of the fourth graders in his study and found higher MLTU and

C/TU but not MLC in the expository essays. In a large study with Hebrew L1 writers (fourth, seventh, and eleventh graders, and graduate students), Ravid (2004) reported higher MLC in expo-argumentative essays than that in narrative essays for basically all the grade levels.

Not much investigation has gone into other language production features in the L1 writing studies comparing narration and argumentation, exposition, or expo-argumentation. Regarding the amount of text generated, San Jose (1972) found the fourth graders in his study produced longer essays on the narrative task than those on the expository and the argumentative tasks. Similarly, Beers and Nagy (2009) reported significantly longer texts generated for narration than those generated for argumentation by seventh and eighth graders. For lexical complexity, Ravid (2004), examining Hebrew L1 writing, found greater lexical density as measured by number of content words per clause for expo-argumentative essays than that for narrative essays written by seventh and eleventh graders, but not by fourth graders. No L1 study has examined the accuracy of language production; but Pringle and Freedman (1979) found that seventh and eighth graders demonstrated much weaker control over rhetorical features and written language conventions in their argumentative essays than that in their narrative essays. Based on the L1 studies, it appears that young L1 writers can write longer and rhetorically more appropriate essays for narrative tasks, but that L1 writers across grade levels produce syntactically and lexically more complex language for expository, expo-argumentative, and argumentative tasks in comparison to narrative tasks.

As far as L2 writing studies are concerned, there is a paucity of studies comparing language production features of narrative essays with those of the other rhetorical tasks. Only one study is relevant in the discussion here. Lu (2011) compared SC in narrative and argumentative writing by college-level Chinese EFL writers who were English-major students.

The author found significantly higher SC in untimed argumentative essays as well as timed and untimed argumentative essays taken together for 13 of the 14 SC measures automated in the author's computational tool–L2 Syntactic Complexity Analyzer.

In sum, by examining actual student writing, the above studies show significantly greater syntactic complexity and possibly greater lexical complexity as well in expository and argumentative discourse than those in narrative discourse.

2.1.2.2.2 Exposition or expo-argumentation vs. argumentation

Few studies have compared language production features of expository or expo-argumentative and argumentative writing, and they all examined college-level writing, but the findings for both L1 and L2 writing do not seem to point to any clear direction, and topic again turns out to be an important factor to consider in a comparison of linguistic performance across these rhetorical tasks.

Park (1988) and Reid (1990) both conducted some linguistic analyses of selected essays, including L1 and L2 essays, from Carson et al. (1985), with Park only examining the farming expository and the outer space argumentative essays and Reid examining essays from all four tasks. In terms of the length of essays, Reid found both L1 and L2 writers produced significantly longer texts for the expository tasks than those for the argumentative tasks; however, Park found the L1 writers in his study sample wrote significantly longer essays for the outer space argumentative task than for the farming expository task, and Park did not find any difference in text length of the essays produced by the L2 writers in his sample. The contradictory findings about text length reported in the two studies can probably be best explained by variations introduced by topic since Park only used data for the two prompts whereas Reid used and combined data for all the four prompts. Regarding syntactic complexity, neither Park nor Reid found differences in SC in either L1 or L2 essays of the two modes with SC measured by MLTU and ratio of free modifiers in Park and by mean sentence length, percentage of short sentences, percentage of complex sentences, and percentage of passive-voice verbs in Reid. Further, Reid also studied lexical complexity features, and she reported use of words with significantly greater average length in the expository writing of L1 writers than that in their argumentative writing, a result not borne out for the L2 writers though, but use of significantly more content words in the argumentative writing than those in the expository writing for both L1 and L2 writers. Since different topics were used in these linguistic-analysis studies, it is unknown whether the findings about lexical and syntactic complexity are due to topic or due to rhetorical task.

For a comparison of language production features of expo-argumentative and argumentative essays, Greenberg (1981) examined text length, SC features, and linguistic accuracy in the writing of college freshmen who were mostly L1 writers of English. He found some minor significant results. The writers produced significantly more words and more T-units for the argumentative tasks. But in syntax, the expo-argumentative essays were significantly more complex than the argumentative ones, with SC measured by C/TU, MLC, and number of words in the free final modifiers. Further, Greenberg examined linguistic accuracy of the students' writing in the two modes and found no difference in the measures used–sentence control errors and vocabulary errors in the essays. Greenberg's study has a relatively better control over topic, since the expo-argumentative and argumentative tasks share the same topics and only two topics were used. However, the findings of the study are largely based on L1 writing; whether and how L2 writers' linguistic performance in the two rhetorical tasks differs is unknown.

To sum up, with the few studies, comparisons of language production features of expositary, expo-argumentative, and argumentative essays are far from being conclusive. In such examinations, topic appears to be a variable that plays a big role. Even essays of the same rhetorical task have been found to significantly differ in language production features across topics, as discussed in a later section. Therefore, other analytical dimensions within the rhetorical tasks to group writing topics seem necessary to further explore prompt differences. In actual studies, control over topic as in Greenberg (1981) is much needed if the real effect of rhetorical task is to be teased out.

2.1.3 Predictive power of CAF features on writing scores for different rhetorical tasks

Two L1 studies which investigated syntactic complexity in narrative and argumentative essays also sought to compare how SC was related to writing scores for the two RTs. In general, it was found that SC was a significant predictor of writing quality for argumentative rather than narrative tasks; however, the effect of SC was also found to depend on grade levels and the specific SC measures used. Crowhurst (1980b) reported differential effects of SC (MLTU) on writing scores across the three grade levels she examined: for sixth graders, SC did not have an effect on scores for either narrative or argumentative writing, for tenth and twelfth graders, SC had a significantly positive effect on scores for argumentative essays but not for narratives, and for twelfth graders, SC was found to have a significantly negative effect on scores for narrative essays. Beers and Nagy (2009), a study of writing of seventh and eighth graders, indicates that the writing quality of narrative and argumentative essays is associated with different SC measures: for argumentative essays, MLC was found to have a significantly positive association with scores, C/TU was found to have significantly negative association with scores, and MLTU had no association with scores; in contrast, for narratives, C/TU had a significantly positive correlation with scores, and MLC and MLTU were not correlated with writing quality. Further, in an L2 study, Spaan (1993) found MLTU a significant predictor of writing scores for the adult ESL writers' impersonal argumentative essays, but not their personal expository essays.

Besides syntactic complexity, not much has been studied about whether linguistic accuracy of essays or writing fluency predicts writing scores for the different rhetorical tasks differently. Studies that examine the predictive power of CAF on writing scores for the different rhetorical tasks using multiple regression are certainly much needed.

2.2 Topic Familiarity, Cognitive Complexity, and Writing Performance

As pointed out earlier, in addition to rhetorical task, topic seems to bring great variations into writing performance. Even within the same RT, topic has been found to have a great effect on writing scores (e.g., Calman, 1986; Clachar, 1999; Gabrielson, Gordon, & Engelhard, 1995; Hamp-Lyons & Mathias, 1994; Tedick, 1990) and on language production features such as text length (Nold & Freedman, 1977; Tedick, 1990; Yang, 2009), syntactic complexity (Crowhurst & Piche, 1979; Tedick, 1990; Yang, Lu, & Weigle, 2012), lexical complexity (Reynolds, 2002; Yang & Weigle, 2011), and grammatical accuracy (Clachar, 1999). In these and other studies examining topic effect on writing performance, topics about different subject matters are typically used; in this sense, it is a topic effect as well as a subject matter effect that has been observed. In some writing studies, topic has been examined with dimensions that regard the cognitive complexity of writing prompts –personal vs. impersonal topics and familiar vs. less familiar topics, both of which tap into topic familiarity.

2.2.1 Topic familiarity and cognitive complexity

Topic familiarity affects cognitive processing in task performance. Based on dual processing theories (Evans, 2010; Evans, 2011; Stanovich, West, & Toplak, 2011), with abundant experience, one builds highly compiled knowledge about the thing being experienced, forming easily and automatically accessible knowledge and rules about the topic content or task. This kind of highly compiled knowledge allows more autonomous and less effortful cognitive processing, whereas lack of knowledge or familiarity for topic content or task places high cognitive loads on one's working memory, leading to slower and more effortful cognitive processing (Evans, 2010; Evans, 2011; Stanovich, West, & Toplak, 2011). Further, topic/content familiarity is one of the cognitive complexity dimensions that the task-based language teaching literature proposes (e.g., Robinson, 2007a; Skehan, 1998), with more knowledge and familiarity seen as cognitively less demanding than lower or no knowledge and familiarity. What follows is a summary of the findings about the dimensions of personal vs. impersonal topics and familiar vs. less familiar impersonal topics which both relate to the topic familiarity cognitive complexity dimension in this study in terms of their relationships with writing performance.

2.2.2 Topic familiarity and second language writing performance

2.2.2.1 Personal vs. impersonal topics

This dimension of writing prompts is mostly called personal vs. impersonal topics in the writing literature, although private vs. public is also used as the term (Hamp-Lyons & Mathias, 1994). Personal vs. impersonal will be adopted here in this review. In Hamp-Lyons and Mathias (1994), personal topics are seen as the ones in which "the writers are asked to say how they feel about something or to use their own experience", and impersonal topics are the ones in which "writers are expected to speak about groups and communities rather than about themselves and/or their families/experiences" (p. 55). Hamp-Lyons and Mathias found impersonal topics cognitively more demanding than personal ones, based on expert judgments on their difficulty as well as theories of human cognitive, moral and affective development (Kohlberg, 1983; Moffett, 1968; Peel, 1971; Piaget, 1972) and L1 writing development (Britton, Burgess, Martin, MacLeod, & Rosen, 1975; Hays, 1983; Wilkinson et al., 1980). In the task-based language teaching literature, personal topics are seen as more familiar than impersonal ones (Skehan, 1998), thus cognitively less demanding.

The effects of personal or impersonal topics on writing quality have been studied by Spaan (1993), Hamp-Lyons and Mathias (1994), and Yu (2007) who all examined this variable with MELAB essay data, and by Hinkel (2002) who also examined ESL writing. Among these studies, Hamp-Lyons and Mathias, Hinkel, and Yu all reported significantly higher scores for essays on impersonal topics than those on personal topics. Spaan, on the other hand, did not find a significant difference in scores for the two topic types. In Hamp-Lyons and Mathias' investigation, the authors reported a significant interaction effect between the personal vs. impersonal dimension and the exposition vs. argumentation dimension, with impersonal ones receiving the highest scores, although the main effect of RT was also significant. This study illustrates the importance of this topic dimension, when it is studied together with RT. For Spaan's inquiry, since only two topics were used for each of the personal and impersonal types and one of the impersonal topics turned out to be particularly difficult, the finding of this study may be less generalizable.

As for the effect of personal or impersonal topics on language production features of ESL writing, Spaan (1993) found that personal topics invited longer texts and linguistically more accurate texts as measured by number of error-free T-units but impersonal ones elicited lexically more complex language as measured by the percentage of words with three or more syllables, and there seemed to be no difference in SC as measured by MLTU and in lexical diversity as measured by type-token ratio; no statistical testing was conducted in this study. Yu (2007; 2010) also found that impersonal topics elicited lexically significantly more complex language as measured by vocd D–an adjusted lexical diversity measure. Additionally, Hinkel (2002) reported more native-like language features in ESL essays on impersonal topics than those on personal topics. It appears that, in general, impersonal topics can invite linguistically more complex language production.

Although this dimension of personal vs. impersonal topics is found to be important in affecting writing performance, it should be noted that such a distinction is not without problems. In particular, studies of stance invitation - taking a personal stance or impersonal stance in response to an essay prompt, have found that writers do not always follow the stance invited (e.g., Greenberg, 1981; Hoetker & Brossell, 1989). For instance, Hoetker and Brossell (1989) reported that as high as 40% of the essays written for impersonally addressed prompts were written actually in the first person, although there were a significantly higher percentage of essays written in the first person for personally addressed prompts than those for impersonally addressed ones. The personal and impersonal distinction may not be a black and white dichotomy as Hamp-Lyons and Mathias (1994) defines, but rather a continuum, when we actually examine writers' written responses. It is plausible that although for personal topics writers would mostly write about their own experiences, feelings, preferences, and so on in relation to the subject matters, for impersonal topics, the amount of direct and explicit knowledge writers have built from experience in relation to a topic is likely to determine whether and to what extent writers can approach an impersonal topic personally. With at least some amount of experiential knowledge, writers can possibly approach an impersonal topic personally. Lack of experiential knowledge, on the other hand, makes it almost impossible for writers to write from their own experiences in relation to the subject matter. In this manner, impersonal topics are probably better seen as having two levels depending on the amount of experiential knowledge writers have on a topic. From here, topic familiarity can be further divided into impersonal familiar and impersonal less familiar topics, although very few writing studies have examined this dimension.

2.2.2.2 Impersonal familiar vs. impersonal less familiar topics

Only two previous writing studies are relevant here for a comparison of impersonal familiar and impersonal less familiar topics. Tedick (1990), examining adult ESL writing, found that essays on a more familiar impersonal topic received significantly higher scores than a less familiar impersonal topic. Further, Tedick reported significantly higher overall syntactic complexity as measured by mean length of T-unit and significantly greater text length produced for the more familiar topic. Yu (2007; 2010), also studying adult ESL writing, found significantly higher lexical diversity for essays on impersonal-familiar topics than that for essays on impersonal-less familiar topics. These studies suggest potentially higher scores and higher linguistic complexity for essays on impersonal familiar topics than those for essays on impersonal-less familiar topics.

2.3 Task Complexity and CAF of L2 Production in the TBLT Literature

In the above section, I summarized the dimensions regarding the cognitive complexity of writing prompts in the L1 and L2 writing literature (i.e., rhetorical task, personal vs. impersonal topics, and impersonal familiar vs. impersonal less familiar topics) and their effects on writing quality and CAF features of writing. In this section, I present the theories and relevant findings in the task-based language teaching (TBLT) literature regarding the effect of the cognitive complexity of tasks on task performance features in the CAF areas, which is an area of great interest in the TBLT literature.

TBLT is an approach to L2 curriculum design, including instruction and assessment, which takes real-world tasks and pedagogical tasks that can promote learners' ability to perform on real-world tasks as the basic units of analysis and curriculum construction (Ellis, 2003; Long & Crookes, 1993; Skehan, 1998). The TBLT framework has its origin in Dewey's (1933) educational concept of "experiential learning", i.e., learning by doing. Some examples of tasks include ordering a pizza on the phone, describing to a friend a movie one watched, and writing an argumentative essay; tasks to include in the syllabus of a course should ideally be based on needs analysis of the types of the tasks that the learners need to be able to perform in the real world (Long & Crookes, 1992). The TBLT approach is in contrast with grammar-translation and functional-notional approaches where linguistic components and language functions and notions are respectively used as the basic units for curriculum design. With tasks as the basic organizing units for TBLT, it does not mean that there is no learning of linguistic components; such learning can occur implicitly or explicitly in the pre-task, during-task, and post-task phases for the pedagogical task that is at the centerpiece (Ellis, 2003).

In the TBLT literature, two somewhat competing hypotheses exist regarding the proposed effects of the cognitive complexity of tasks on CAF features of language production–Robinson's Cognition Hypothesis (Robinson, 2001, 2003, 2005, 2007a, 2010) and Skehan's Trade-off Hypothesis (Skehan, 1992, 1996, 1998; Skehan & Foster, 2001). One purpose of the current study is to test the hypotheses, in the context of L2 writing. These hypotheses have not been formally distinguished for speaking and writing modalities. The majorities of the existing studies examining the hypotheses are studies of spoken tasks, but a handful of them also used written tasks (e.g., Robinson, Ting, & Urwin, 1995; Ellis & Yuan, 2004; Ojima, 2006; Ishikawa, 2007; Kuiken & Vedder, 2007; Kuiken & Vedder, 2008; Ong, & Zhang, 2010; Kormos, 2011). Further, Skehan and Foster (2001) specify modalities as part of task conditions that may have an influence on how the cognitive complexity of tasks affects CAF performance; however, there are no hypotheses as to what the influences may be like, and the authors encourage studies looking into different modalities and revealing findings pertaining to specific modalities. Studies that can test the hypotheses in the writing modality, such as the current one, are certainly much needed. The different nature of speaking and writing modalities, with speaking imposing more real-time pressure for immediate language production and writing allowing more opportunities for planning and editing during task performance, may very likely affect the studied relationships.

2.3.1 CAF in the TBLT literature

As discussed earlier, some writing studies in the L1 and L2 writing literature have examined CAF features of writing affected by factors such as rhetorical task and topic familiarity. Those studies are in general intended to examine the effects as they are, with some having implications for teaching and assessing writing. However, in the TBLT literature, CAF features in language production under different task designs and conditions are given an additional layer of meaning, namely, language development. Complexity, accuracy, and fluency are the three task performance areas that have been selected and often examined in totality, with task designs and conditions being conducive to development of some or all of the areas a point of great interest to researchers.

Skehan (1992, 1996) first suggested these three areas of CAF as the goals in task-based language instruction, as part of the general goal for L2 learners to achieve native-like language performance. The performance areas of accuracy, complexity, and fluency have their weight in both effective communication and language development. As L2 learners progress in their language competence, they should be producing more accurate, fluent, and complex language in their language production. Task-based instruction has the goal of pushing learners towards these performance goals, with a proper sequence of language learning tasks taking into consideration factors such as task complexity. According to Skehan, accuracy relates to “a learner's capacity to handle whatever level of interlanguage complexity s/he has currently attained”, complexity entails restructuring of interlanguage and concerns “the stage and elaboration of the underlying interlanguage system”, and fluency is about “the learner's capacity to mobilize an interlanguage system to communicate meanings in real time” (Skehan, 1996, p. 46). In essence, then, in the TBLT literature, accuracy shows and enables control of a learner’s interlanguage, complexity demonstrates and pushes restructuring and stretching of the interlanguage, and fluency displays and requires a normal speed of accessibility of the interlanguage. The cognitive complexity of tasks is one main factor that is hypothesized to affect these three performance areas, thus playing a role in language performance as well as language development. In task-based language assessment, a concern is also placed on how to elicit L2 learners’ best performance in the three areas so that their interlanguage can properly assessed, based on manipulations of the cognitive complexity of tasks (Skehan, 2001).

2.3.2 The Cognitive complexity of tasks and CAF of L2 production

In the TBLT literature, two somewhat competing hypotheses exist regarding the relationship between the cognitive complexity of tasks and language performance in the areas of CAF–Robinson’s Cognition Hypothesis (Robinson, 2001, 2003, 2005, 2007a, 2010) and Skehan’s Trade-off Hypothesis (Skehan, 1992, 1996, 1998; Skehan & Foster, 2001). There are different categories of cognitive complexity of tasks in the two scholars’ frameworks, and the hypotheses proposed converge and diverge depending on the categories.

Table 2.1 below lists the cognitive complexity factors from the two frameworks. As can be seen, the operationalizations of cognitive complexity in the two schemes are rather different. Robinson (2007a) provides the latest version of his framework; the earliest version is in Robinson (2001). In Robinson's Triadic Componential Framework, task complexity (cognitive factors) is one of the main three categories, along with task condition (i.e., interactive factors including participation variables and participant variables) and task difficulty (i.e., learner factors including ability variables and affective variables). Robinson proposes task complexity as the single dimension to use to sequence pedagogic tasks from simple to complex. The task complexity (cognitive factors) category is further divided into resource-directing and resource-dispersing subcategories; Robinson makes different predictions on the effects of cognitive complexity on CAF for dimensions in the two subcategories, which will be summarized later. Factors in the resource-directing category make cognitive/conceptual demands on learners and thus direct learners' attentional and memory resources to form-function mappings in more complex tasks, and in the latest version (Robinson, 2007a), the category includes six factors such as +/− here and now and +/− causal reasoning. The +/− sign in Robinson's framework means yes or no but also applies to the amount of a specific cognitive function required (Robinson, 2001). For instance, +/− causal reasoning means causal reasoning is involved or not involved in a task, as well as how much causal reasoning is involved, and a task with + causal reasoning would promote learners' attention to forms although it is cognitively more complex. Dimensions in the resource-dispersing category make performative/procedural demands on learners and thus take away attentional and memory resources available for focusing on forms in more complex tasks, and in the latest version (Robinson, 2007a), the category includes six factors such as +/− prior knowledge and +/− task structure. For example, a task where learners lacked prior knowledge is seen cognitively more complex, and it would inhabit the learners’ ability to attend to forms.

Table 2.1

Operationalizations of the Cognitive Complexity of Tasks in TBLT Literature

Robinson (2007a)	Skehan (1998)
Task complexity (Cognitive factors)	2 *Cognitive complexity*
(a) Resource-directing variables	Cognitive familiarity
making cognitive/conceptual demands	- familiarity of topic and its predictability
+/− here and now	- familiarity of discourse genre
+/− few elements	- familiarity of task
+/− spatial reasoning	Cognitive processing
+/− causal reasoning	- information organization
+/− intentional reasoning	- amount of 'computation' **[-** transformation or manipulation of information]
+/− perspective-taking	- clarity and sufficiency of information given
	- information type
(b) Resource-dispersing variables	
making performative/procedural demands	
+/− planning time	
+/− prior knowledge	
+/− single task	
+/− task structure	
+/− few steps	
+/− independency of steps	

Adopted from Robinson (2007a, pp. 15-16) and Skehan (1998, p. 99)

Skehan’s (1992; 1996; 1998) framework of task difficulty/complexity includes cognitive complexity, as well as code complexity (i.e., the difficulty of the language demanded to complete a task) and communicative stress (i.e., performance conditions affecting processing and impacting communication pressure). The cognitive complexity category includes two subcategories – cognitive familiarity and cognitive processing, with the former involving the level of “pre-packaged solutions” available and the latter referring to the “amount of on-line computation” required to work on task content (Skehan, 1996, p. 52).

Comparing the cognitive complexity dimensions in the two classification schemes, the only overlapping ones seem to be the +/− prior knowledge factor in Robinson (2007a) and the cognitive familiarity factor in Skehan (1998), as well as the +/− task structure factor in Robinson and the information organization factor in Skehan with both related to the clarity, predictability, and availability of information structure of tasks. The two scholars operationalize cognitive complexity in rather different ways, but it also seems that the cognitive processing category in Skehan, meaning "amount of on-line computation" required to work on task content (Skehan, 1996, p. 52), can cover many of the factors in both the resource-directing and resource-dispersing categories in Robinson's, since arguably most of the factors in Robinson's require more thinking and on-line 'computation' of the task content if they are made more complex. Researchers who have used Skehan's entire framework of task difficulty/complexity for task design have not found the framework easy to use in designing tasks (Iwashita, McNamara, & Elder, 2001; Norris et al., 1998). No research seems to have used Robinson's entire framework of task complexity for task design.

With regard to predictions on the effects of the cognitive complexity of tasks on CAF, I concur with Robinson and Gilabert's (2007) conclusion that the two hypotheses have the same predictions for the factors in the resource-dispersing category in Robinson's framework, with both predicting lower accuracy and complexity in more complex tasks in this category, but that where the two hypotheses differ is over "the claims described … for the beneficial effects on accuracy and complexity of increasing the resource-directing dimensions of tasks" (p. 167).

Specifically, Robinson's Cognition Hypothesis predicts greater accuracy and complexity and lower fluency for complex tasks along the resource-directing dimensions, and lower accuracy and complexity for complex tasks along the resource-dispersing dimensions. The

Cognition Hypothesis states, “complexity on the former resource-directing dimensions of task demands promotes attention to form-function/concept mappings, thereby leading to interlanguage development, and on the latter resource-dispersing dimensions it promotes increasing automatic access to current linguistic resources." (Robinson, 2010, p. 247). The performance effects of greater attention to form-function/concept mappings in complex conditions along the resource-directing dimensions are that L2 learners shall produce more accurate and complex, but less fluent language. Robinson’s argument for the effects of the dimensions in this category on syntactic complexity is largely based on the notion that “greater structural complexity tends to accompany greater functional complexity in syntax” (Givon, 1985, p. 1021; see also Givon, 1989; Sato, 1988; 1990), as well as staged development of conceptual/cognitive abilities with their attendant linguistic codes in childhood (e.g., Bartsch & Wellman 1995; Cromer 1974) and in adult naturalistic L2 acquisition (e.g., Becker & Carroll 1997; Perdue 1993). The argument for more accurate language production in more complex conditions along the resource-directing dimensions is primarily based on the assumptions that greater conceptual and functional demands push greater attention to accuracy of form (e.g., Hulstijn, 1989; Tarone, 1985) and that the greater demands can also potentially make learners notice how their L1 and L2 probably grammaticize conceptual notions differently (Talmy 2000; von Stutterheim 1991). Robinson’s theorizing about the effects of task complexity on CAF is also based on a multiple-resource model of attention (e.g., Navon, 1989; Wickens, 1989), which means that content and form may not always be competing for scarce attentional resources.

For the resource-dispersing dimensions, increased complexity for dimensions in this category, according to Robinson (2001), do not promote interlanguage development, but rather promote “consolidation and fast real-time access to existing interlanguage resources" (Robinson, 2010, p. 252). Performing more complex versions of tasks along these dimensions “leads to a *depletion* of attentional and memory resources” (Robinson, 2001, p. 308, emphasis in original), and results in lower accuracy and complexity in learners’ L2 production (Robinson & Gilabert, 2007). In addition to the claims about the separate performance effects for dimensions in the two categories, Robinson also predicts the synergetic effect for the two categories, arguing that the positive effects on accuracy and complexity of increased task complexity along the resource-directing dimensions (e.g., + causal reasoning) will be stronger if the complexity along the resource-dispersing dimensions is reduced, i.e., if simple versions in this category are used (e.g., + prior knowledge) (Robinson, 2005; Robinson & Gilabert, 2007). It should also be pointed out here that the above claims about the effects of task complexity along the resource-directing dimensions on CAF apply to monologic tasks, and that for interactive tasks, due to increased number of comprehension checks and clarification requests using phrasal and one-word responses, syntactic complexity of language production is not likely to increase in more complex interactive tasks (Robinson, 2001).

Robinson thus makes different predictions on the effects of the cognitive complexity of tasks on CAF for dimensions in the two different categories he has in his framework. Skehan, on the other hand, does not have such a distinction in the predicted effects. Rather, the Trade-off Hypothesis (Skehan, 1992, 1998; Skehan & Foster, 2001) predicts that increased cognitive complexity of tasks leads to a competition among accuracy, complexity, and fluency, with an increase in one area often at the sacrifice of another area. One main assumption that the Trade-off Hypothesis rests on regards a limited-capacity attentional system, for which meaning and form in language performance compete for the limited attentional resources (VanPatten, 1990; 1994). VanPatten (1990; 1994) found that in communicative situations, language learners prioritize meaning over form. Thus, a tension exists between meaning (fluency) and form (accuracy and complexity). In addition, according to Skehan, accuracy and complexity also enter into competition depending on whether L2 learners choose to be conservative by using controlled interlanguage, thus attending to accuracy, or choose to take risks in extending and stretching the interlanguage, thereby attending to complexity.

For the operationalizations of cognitive complexity that the current study examines–rhetorical task and topic familiarity, the two hypotheses would have the same predictions regarding the effects on CAF from topic familiarity which is similar to +/− prior knowledge in the resource-dispersing dimensions in Robinson's framework and the familiarity of topic dimension in Skehan's scheme. Both would predict lower accuracy and complexity and possibly lower fluency as well when writers have lower topic familiarity. The two hypotheses however would have different predictions for the effects of rhetorical task on CAF, since the different cognitive demands inherent in the rhetorical tasks seem to largely fall into Robinson's resource-directing dimensions in which there are reasoning factors. The reasoning factors in Robinson's latest framework (Robinson, 2007a) relate to very specific concepts/ functions, i.e., +/− causal reasoning, +/− intentional reasoning, and +/− spatial reasoning, although the earlier version of the framework (Robinson, 2001; 2003; 2005) only has one general reasoning – +/− reasoning. The reasoning demanded in narrative, expository, expo-argumentative, and argumentative tasks is rather in general terms with narrative and expository tasks involving basically no reasoning, expo-argumentative tasks demanding some reasoning, and argumentative tasks requiring much more reasoning. Further, in comparison to narrative tasks which only primarily only needs recalling, expository tasks require generalizing. Then based on Robinson's Cognition Hypothesis, the predicted effects of rhetorical task on CAF are that writers will produce language of higher accuracy and complexity but lower fluency when they perform on cognitively more demanding rhetorical tasks, while Skehan’s Trade-off Hypothesis would not predict such beneficial effects on accuracy and complexity.

As for empirical studies in the TBLT literature examining the effects of the cognitive complexity of tasks on CAF, there are a good number of them; however, the vast majority of the studies used speaking tasks rather than writing tasks, and the factors in Robinson’s and Skehan’s frameworks that have been studied are limited to a few, with availability of planning time and types of planning given the most attention (See Ellis, 2009 for a review) and with +/− here and now, +/− task structure, and +/− reasoning receiving some attention. Few studies have examined the effect of familiar vs. less familiar topics on task performance, and basically no study has examined reasoning in the broad sense of narration, exposition, expo-argumentation, and argumentation.

Foster and Skehan (1996) and Skehan and Foster (1997) examined the familiar vs. less familiar dimension through speaking tasks of three different types, namely, a personal task (i.e., task on personal topics), a narrative task based on picture strips, and a decision-making task based on scenarios, with the personal task seen as having content the most familiar to the ESL speakers and the decision-making task seen as having content the least familiar and also the most unpredictable to the participants. The two studies used different sets of tasks, and both also examined the effect of planning, with Foster and Skehan having three planning conditions–no planning time, unguided planning, and guided planning and Skehan and Foster having two planning conditions–no planning time and 10 minutes of planning. As for the effect of information familiarity on fluency, Foster and Skehan found the personal task, the most familiar one, generated the most fluent speech as measured by the number of pauses and the total amount of silence in both the no planning and unguided planning conditions, and Skehan and Foster found the narrative task elicited the most fluent speech as measured by the number of pauses in both no planning and planning conditions; yet both studies reported lower fluency in performance on decision-making task, the task with the least familiar information. Opposite patterns were observed for syntactic complexity as measured by the number of clauses per communication unit in both studies: the personal task in Foster and Skehan elicited the lowest SC in both planning conditions and the narrative task in Skehan and Foster generated the lowest SC in both planning conditions, and the decision-making tasks, the tasks with the least familiar information, elicited higher SC in comparison to the personal tasks in both studies. For the effect of content familiarity on accuracy, it seems that the narrative tasks in both studies, the tasks the authors regarded as having medium content familiarity to the participants, elicited the least accurate language, but the pattern was observed in different planning conditions in the two studies.

To interpret Foster and Skehan's (1996) and Skehan and Foster's (1997) findings on the effects of content familiarity on CAF, we have to bear two factors in mind. First, the tasks of the same types used in the two studies are not quite comparable in difficulty, as Skehan (1998) points out. Second, the studies seem to have confounded content familiarity and rhetorical task, since the personal tasks are in fact personal expository tasks, the narrative tasks are impersonal narration since they are based on picture stories that have not been seen or heard before, and the decision-making tasks are similar to impersonal argumentative tasks. But in general, the findings of the two studies corroborate on the findings from the writing studies presented earlier, showing higher fluency in personal expository tasks (Spaan, 1993) and in narrative tasks (San Jose, 1972;

Beers & Nagy, 2009) and higher syntactic complexity in argumentative tasks (e.g., Crowhurst & Piche, 1979; Lu, 2011).

2.4 Measures of CAF in the TBLT and the Writing Literature

In both the TBLT literature and the L2 writing literature, the constructs of CAF are examined and measured. In the TBLT literature, as laid out earlier, the interest is predominately within the effects of task complexity on CAF in language production. In the L2 writing literature, task effects on CAF have not been much of a concern, but a number of L2 writing studies have investigated how CAF in writing relate to L2 proficiency levels or L2 writing quality (Polio, 2001; Ortega, 2003; Wolfe-Quintero, Inagaki, & Kim, 1998). In this section, the measures used for CAF in the two research areas are summarized and compared. More importantly, the underlying constructs of CAF are considered and discussed for studies that intend to examine the effects of the cognitive complexity of tasks on CAF. The summary of the measures used in the TBLT literature is based on my own literature review, and the summary of the measures used in the L2 writing literature is based on Polio (2001).

In the TBLT literature, measures of CAF have been used in studies examining the effects of the cognitive complexity of tasks on task performance in the three performance areas, with the complexity factors of +/− planning time (See Ellis, 2009 for a review), +/− here and now (e.g. Ishikawa, 2007), +/− reasoning (Robinson, 2007b), and +/− task structure (e.g., Kormos, 2011) studied the most. Through an extensive review of the TBLT literature, I identified a total of 40 studies that examined the effects of the cognitive complexity of tasks on CAF. Most of the studies examined all the three performance areas of CAF, but some studies examined only two of them. For the performance area of complexity, both syntactic complexity and lexical complexity have been studied, with some studies examining one of them and others looking into both. Many studies used more than one measure for a performance area. Table 2.2 outlines, for each construct, the total number of studies that have examined it, the measures used with percentage information for how many studies out of the total number of studies used that measure, and one representative study for each measure. As can be generally observed in the table, there are many different measures used for each construct, but certain measures have been more frequently used.

For measures of CAF used in the L2 writing literature, Table 2.3 shows the types of measures used for each construct, based on Polio (2001) which unfortunately does not provide frequency information for each. What follows is an examination of each construct and how it is measured in the two research areas.

Table 2.2
Measures for Accuracy, Complexity, and Fluency Used in TBLT Studies

Accuracy	**Syntactic Complexity**
39 studies in total	**33 studies in total**
- percentage of error-free clauses (Skehan & Foster, 1997) (44%) - percentage of error-free production units (T-units mostly) (Ishikawa, 2007) (23%) - percentage of target-like use of articles (Crookes, 1989) (18%) - percentage of target-like use of other specific morphological or grammatical forms (Ortega, 1999) (15%) - normalized total number of errors (Sangarun, 2005) (15%)	- amount of subordination (mostly number of clauses per production unit - T-unit, c-unit, or AS-unit) (Skehan & Foster, 1997) (73%) - S-nodes (Bardovi-Harlig, 1992) per production unit (T-unit mostly) (Gilabert, 2005) (36%) - mean length of production unit - AS-unit, c-unit, T-unit, or utterance (Foster & Tavakoli, 2009) (24%)
less than 10% of the 39 studies used:	***less than 10% of the 33 studies used:***
- percentage of verb-related errors (Ellis & Yuan, 2004) - amount of error repairs (Gilabert, 2007) - percentage of word order errors (Mehnert, 1998) - percentage of lexical choice errors (Mehnert, 1998) - mean length of error-free clauses (Skehan & Foster, 2005)	- total number of multi-propositional utterances (Sato, 1990) (Robinson, 1995) - total number of different verb forms used (Ellis & Yuan, 2004) - number of propositions per utterance (Ortega, 1995) - number of words before main verbs (Kormos, 2011) - number of modifiers per head noun (Kormos, 2011) - total number of coordinated verb phrases (Wendel, 1997) - total number of passive voice used (Wendel, 1997)

Lexical Complexity	Fluency
25 studies in total - Lexical diversity measures (type-token ratio, adjusted type-token ratio) (Ellis and Yuan (2004) (68%) - Lexical density measures (lexical words/ total words, lexical words/ functional words) (Robinson, 1995) (36%) ***less than 10% of the 25 studies used:*** - Lexical sophistication (lexical frequency profile, Laufer & Nation, 1995) (Kormos, 2011) - Range of verb forms (Crookes, 1989)	**32 studies in total** - Breakdown fluency (e.g., total number of pauses greater than .4 of a second, total amount of silence) (Foster & Tavakoli, 2009) (53%) - Repair fluency (e.g., number of word replacement, false start, reformulation, repetition) (Foster & Skehan, 1999) (44%) - Speech rate (e.g., syllables per second, words per minute) (Gilabert, 2005) (38%) ***less than 10% of the 32 studies used:*** - Total number of words (Ishikawa, 2007) - Writing rate (words per minute) (Ong & Zhang, 2010)

Table 2.3
Measures for Accuracy, Complexity, and Fluency Used in Writing Studies

	Accuracy	
Holistic scales	Jacobs scale	Hedgcock & Lefkowitz (1992)
	Other scales	Tarone et al. (1993)
	Binary classification	Devine et al. (1993)
Error-free units	Error-free clauses/total clauses	Ishikawa (1995)
	Error-free T-units/total T-units	Polio et al. (1998)
	Words in EFT/total words	Polio et al. (1998)
Number of errors	Without classification	Carlisle (1989)
	With classification	Frantzen (1995)
Qualitative analysis		Shaw & Liu (1998)
	Syntactic complexity	
Average length of a structure	Words per t-unit	Cooper (1981)
Frequency of a structure	Passive	Kameen (1979)
	Dependent clause	Homburg (1984)
Complexity ratio	Clauses/t-unit	Bardovi-Harlig & Bofman (1989)
	Coordination index	Bardovi-Harlig (1992)
Qualitative analysis	Syntactic profile	Coombs (1986)
	Lexical complexity	
Lexical individuality/originality	Ratio of tokens unique to a writer/number of tokens	Laufer (1991)
Lexical sophistication	Advanced token/total tokens	Engber (1995)
	Lexical Frequency Profile	Laufer & Nation (1995)
Lexical variation/diversity	Different types/total tokens	Frantzen (1995)
Lexical density	Lexical words/total words	Laufer (1991)

	Fluency	
Holistic scales		Tarone et al. (1993)
Amount of production	Words	Henry (1996)
	T-units	Ishikawa (1995)
	Clauses	Robb et al. (1986)

Adapted from Polio (2001) Table 7.2 (p. 94), Table 7.3 (p. 96), Table 7.4 (p. 99), & Table 7.8 (p. 106)

2.4.1 Accuracy measures

Linguistic accuracy, as Polio (2001) defines it, is “a broad term that generally has to do with the absence of errors” (p. 94). Errors broadly include morphological and grammatical errors, as well as errors in word choice, spelling and punctuation. What errors are counted varies in studies. To measure the accuracy construct, it may be ideal to account for all types of errors in a weighted manner as certain errors are more obtrusive to meanings than others, but as Tables 2.2 shows, general measures that do not take into account the seriousness of the errors, such as percentage of error-free clauses and percentage of error-free T-units, have been commonly used in TBLT studies, as well as measures that capture certain specific kinds of errors only, such as percentage of target-like use of articles. Comparing the accuracy measures in Table 2.2 and in Table 2.3, it can be observed that they share the general measures of percentage of error-free clauses (or T-units) and total number of errors, but accuracy measures for specific linguistic forms such as percentage of target-like use of articles seem more common in TBLT studies and measures through holistic scales and qualitative analysis are found in writing studies only.

2.4.2 Syntactic complexity measures

Syntactic complexity (SC) regards the complexity of constructions used in sentences. The construct is traditionally associated with complex sentences in which there is more than one clause (See Ravid & Berman, 2010). The notion of clauses, in grammar theories (Cristofaro, 2003; Givon, 2008; Halliday & Matthiessen, 2004; Langacker, 2008), includes both finite clauses and nonfinite clauses. This level of SC is at the clausal level, involving non-simple sentences and more than one verb with coordinated verbs excluded. In addition to complex sentences, phrasal complexity (particularly, complexity of noun phrases) has been proposed to represent syntactic complexity at another level–the sub-clausal level (Norris & Ortega, 2009), a view that can find its support in L1 and L2 syntactic development studies (e.g., Cooper, 1976; Crossley et al., 2011; Hunt, 1965; Lu, 2011; Ravid & Berman, 2010) and studies of discourse analysis of spoken and written texts in different genres (e.g., Biber, 2006; Biber, Gray, & Poonpon, 2011; Ravid & Berman, 2010). In a nutshell, as Norris and Ortega (2009) proposes, syntactic complexity is a multidimensional construct, with representations at the global level, clausal subordination level, and sub-clausal level. This multidimensional nature of the SC construct calls for measures at these different levels.

As Tables 2.2 and 2.3 show, subordination measures, particularly clauses per production unit (e.g., T-unit), are commonly used in TBLT and writing studies, so are measures tapping into global complexity–mean length of a main production unit (e.g., T-unit). However, measures examining noun-phrase complexity have been rarely used in TBLT studies, with the exception of Kormos (2011), and have also been seldom used in writing studies, which seems problematic since noun-phrase complexity has been found to feature in formal expository and argumentative discourse (e.g., Biber, Gray, & Poonpon, 2011; Hunt, 1965; Ravid & Berman, 2010).

Table 2.2 and Table 2.3 show that the main types of SC measures used in the two areas largely overlap. It should however be noted that subordination measures such as clauses per T-unit (C/TU) do not mean the same in the two research areas. As reviewed earlier, in the writing literature, following Hunt (1965), the word “clause” only refers to finite clauses with nonfinite clauses excluded from the picture, and thus clauses per T-unit in the writing literature in fact means finite clauses per T-unit, measuring the amount of finite subordination only. In the TBLT literature though, the majority of the studies have not specified how clauses were defined and counted, even though the earlier pioneers and leading scholars for those studies, with Foster and Skehan's (1996) and Robinson's (1995) work as the most representative, counted both finite- and non-finite clauses as clauses in the subordination measures they used. Only three of the 33 TBLT studies explicitly stated that only finite clauses were counted as clauses (Ishikawa, 2007; Mehnert, 1998; Wigglesworth, 1997).

Another difference in SC measures in TBLT studies and in writing studies is that clausal subordination is the most commonly used SC measure used in TBLT, as can be seen in Table 2.2, whereas mean length of T-unit–a global SC measure, is the most commonly used SC measure in L2 writing studies (Ortega, 2003). Additionally, mean length of sentence–another global SC measure taking into account clausal coordination as well, has been only occasionally used in writing studies; yet, in a recent investigation of the predictive values of 14 SC measures on writing scores (Yang, Lu, & Weigle, 2012), mean length of sentence was revealed as a significant predictor of writing scores for essays on both of the argumentative tasks examined.

2.4.3 Lexical complexity measures

Probably even more than syntactic complexity, lexical complexity has been seen as a multidimensional global construct. Lexical richness is also the term used in the literature (Read, 2000). This global construct is used to refer to a set of lexical features that make texts vary in lexical diversity, lexical density, and lexical sophistication/ rareness (Polio, 2001; Read, 2000; Wolfe-Quintero et al., 1998). These lexical features are probably best seen as sub-constructs of lexical complexity. Specifically, according to Read (2000), lexical diversity or variation refers to the number of different words used in relation to the total number of words in a text, lexical density has to do with the proportion of lexical words in a text to that of grammatical words, and lexical sophistication or rareness regards the number of sophisticated/ rarer words or word families used out of the total number of words or word families in a text.

As can be seen in Tables 2.2 and 2.3, all the three sub-constructs of lexical complexity have been measured in TBLT and in writing studies. In TBLT studies, lexical diversity measures have been the most popular, and only one study has examined lexical sophistication. However, both lexical diversity and lexical sophistication have been found to be revealing indices of lexical richness in both first and second language developmental studies (e.g. Daller, van Hout, & Treffers-Daller, 2003; Laufer & Nation, 1995; Malvern et al., 2004; van Hout & Vermeer, 2007) and have thus been recommended for use alongside each other. Further, measures for both have been found to positively correlate with writing scores (e.g., Yang & Weigle, 2011; Yu, 2010). In actual L2 studies, the two sub-constructs have been rarely examined together. On the other hand, in comparison to lexical diversity and lexical sophistication, lexical density has not been found to differentiate language proficiency levels (e.g., Engber, 1995; Linnarud, 1986; Lu, 2012). However, as Ravid's (2004) study reveals, lexical density is sensitive to rhetorical task, with expository and argumentative essays significantly lexically denser than narrative essays.

When we measure syntactic complexity and lexical complexity in studies examining the effects of the cognitive complexity of tasks on linguistic complexity, it appears important to understand which sub-constructs of syntactic complexity and which sub-constructs of lexical complexity are affected when the cognitive complexity of tasks is varied on a certain dimension such as rhetorical task and topic familiarity. In this regard, different sub-contructs of linguistic complexity may function differently.

2.4.4 Fluency measures

Fluency in speaking and in writing do not mean the same thing, and thus is measured quite differently in the two modalities, as can be observed in the difference in the fluency measures in Table 2.2 where most of the TBLT studies used speaking tasks and those in Table 2.3 for writing studies. The picture for fluency measures for speaking tasks is more complicated, as the construct takes into account pausing, false-starts, speech rate, and several other factors.

In the context of writing, Wolfe-Quintero, Inagaki, and Kim (1998) defines fluency as follows:

> In our view, fluency means that more words and more structures are accessed in a limited time, whereas a lack of fluency means that only a few words or structures are accessed. Learners who have the same number of productive vocabulary items or productive structures may retrieve them with differing degrees of efficiency. Fluency is not a measure of how sophisticated or accurate the words or structures are, but a measure of the sheer number of words or structure units a writer is able to include in their writing within a particular period of time. (p. 25)

This view of fluency is reflected in the commonly used fluency measures in the L2 writing studies: total number of words, T-units or clauses produced. However, as Polio (2001) points out, total number of T-units or clauses is a problematic measure since learners who produce longer T-units or clauses would be penalized with such a measure.

In addition to the view of fluency meaning the amount of text and structures generated in a given amount of time, in writing studies, fluency is occasionally seen to mean how native-like the language production is, as in Tarone et al.'s (1993) rating scale for fluency–"nativeness, standardness, length, ease of reading, idiomaticity" (p. 170). In the TBLT literature, such a "nativeness" view has not been adopted for measuring fluency, and quantifiable text features are rather used.

2.5 Gaps in Previous Research

Based on the above review of the L1 and L2 writing literature and the TBLT literature related to the two main cognitive complexity dimensions of rhetorical task and topic familiarity, several research gaps are identified, and the current study aims to fill these gaps. First of all, the current study aims to bridge the two lines of literature that have both examined and/or discussed the cognitive complexity of tasks and CAF variables in an examination of the two cognitive complexity dimensions. In previous research, this has not been done, and the researchers in these two lines of literature do not seem to have cited each other's work. Second, the current study sets out to fill the research gap that there have not been enough writing studies that test the hypotheses made in the TBLT literature about the effects of the cognitive complexity of tasks on CAF features. An examination of the relationships in the context of timed essay writing is even more lacking. Third, no previous research has systematically examined the predictive power of CAF variables on writing quality scores for tasks of different cognitive complexity along the two studied dimensions or has made comparisons across the tasks; the current study has the goal of examining these unexplored relationships.

In addition to the above three main research gaps, the study also aims to fill some gaps in research methodologies in related inquires. First, few previous studies have examined all the rhetorical tasks of narration, exposition, and argumentation, and no previous study has investigated all the topic familiarity tasks of personal-familiar, impersonal-familiar, and impersonal-less familiar tasks. The current study examines all these levels for rhetorical task and topic familiarity and explicitly defines these levels. Further, the study also explores expo-argumentation–a category created in between exposition and argumentation, to address the inconsistencies in the literature in categorizing this type of task. Second, previous writing research has suggested the importance of controlling for the other dimension in an examination of one of the cognitive complexity dimensions and controlling for the subject matter of the tasks; however, these have not always been attended to (e.g., Park, 1988; Reid, 1990; Spaan, 1993). The current study aims to circumvent the previous limitations by exercising a good control over these design features. Finally, no previous study has examined all the main sub-constructs of lexical complexity as laid out in Read (2000) and of syntactic complexity as proposed in Norris and Ortega (2009); however, the different sub-constructs may function differently in relation to task types and writing scores. The current study attempts to fill the methodological gaps by investigating and measuring all the main sub-constructs of lexical and syntactic complexity.

2.6 Research Questions and their Hypotheses

With an identification of the above research gaps in previous research, the current study pursues six research questions, and they are restated as follows, along with the hypotheses made for the questions. The hypotheses are primarily based on the findings from previous L1 and L2 writing studies that were described in this chapter and are briefly summarized below in brackets following each hypothesis. Since there are no previous studies that have looked into the relationships and have made the comparisons for questions 3 and 6, no hypotheses were made for these two questions.

1. What is the effect of varying the cognitive complexity of independent writing tasks along the rhetorical task dimension on L2 writing scores of college-level ESL writers'?

   H_1: Rhetorical task does not have an effect on L2 writing scores of college-level ESL writers'. [Spaan (1993) and Lim (2009), examining adult-level ESL writing, did not find any difference in scores on different rhetorical tasks.]

2. What is the effect of varying the cognitive complexity of independent writing tasks along the rhetorical task dimension on linguistic complexity, accuracy, and fluency of L2 writing production of college-level ESL writers'?

   H_1: When the cognitive complexity of writing tasks increases along the rhetorical task dimension, linguistic complexity of the L2 production by college-level ESL writers increases, accuracy decreases, and fluency is not affected. [For syntactic complexity, many previous studies found global syntactic complexity of argumentative essays to be significantly higher than that of narrative essays across grade levels (Beers & Nagy, 2007; Crowhurst & Piche, 1979; Crowhurst, 1980a; Lu, 2011; San Jose, 1972). For lexical complexity, Ravid (2004) found significantly higher lexical density in expo-argumentative essays than that in narrative essays, but no previous study has compared the lexical diversity or lexical sophistication of essays on different rhetorical tasks. For accuracy, Pringle and Freedman (1979) reported that younger L1 writers showed weaker control over writing conventions in argumentative essays than that in narrative essays. For fluency, Beers and Nagy (2009) and San Jose (1972) found the narrative essays produced by younger L1 writers to be longer than the expository and argumentative essays they wrote, but Greenberg found the argumentative essays produced by college freshmen to be longer than the expo-argumentative essays they produced, suggesting that adult writers are probably able to produce essays of the same length on different rhetorical tasks.]

3. How do linguistic complexity, accuracy, and fluency of L2 writing production predict L2 writing scores of college-level ESL writers' for independent writing tasks of different

cognitive complexity along the rhetorical task dimension?

4. What is the effect of varying the cognitive complexity of independent writing tasks along the topic familiarity dimension on L2 writing scores of college-level ESL writers'?

   H_1: College-level ESL writers produce the highest L2 writing scores on impersonal familiar tasks, higher than those on personal familiar and impersonal less familiar tasks. [Hamp-Lyons and Mathias (1994), Hinkel (2002), and Yu (2007), all studies of adult L2 writing, found higher scores for essays on impersonal topics than those on personal topics, and Tedick (1990), also a study of adult L2 writing, found higher scores for essays on an impersonal familiar topic than those on an impersonal less familiar topic. These studies suggest the highest writing quality scores on impersonal familiar tasks.]

5. What is the effect of varying the cognitive complexity of independent writing tasks along the topic familiarity dimension on linguistic complexity, accuracy, and fluency of L2 writing production of college-level ESL writers'?

   H_1: When the cognitive complexity of writing tasks increases along the topic familiarity dimension, accuracy and fluency of the L2 production by college-level ESL writers decrease, and the highest linguistic complexity is achieved for impersonal familiar tasks, higher than that for personal familiar and impersonal less familiar tasks. [Spaan (1993) found personal essays to be longer and linguistically more accurate than impersonal essays. Yu (2007; 2010) found higher lexical diversity in essays on impersonal topics than that for personal topics and higher lexical diversity in essays on impersonal familiar topics than that for impersonal less familiar topics, suggesting the highest lexical diversity for impersonal familiar tasks. No previous studies have compared lexical sophistication or lexical density in essays varying in topic familiarity. Tedick (1990) found higher global syntactic complexity in essays on an impersonal familiar topic than that for an impersonal less familiar topic. All these studies examined adult L2 writing.]

6. How do linguistic complexity, accuracy, and fluency of L2 writing production predict L2 writing scores of college-level ESL writers' for independent writing tasks of different cognitive complexity along the topic familiarity dimension?

CHAPTER 3: METHODOLOGY

This chapter outlines the research methods used to answer the research questions, including the participants recruited, the research materials used, recruitment and consent procedures, data collection procedures, essay rating, CAF measures used, and data analysis.

3.1 Participants

A total of 375 Chinese EFL university students in China participated in the study, with approximately 61-64 students writing essays on each of the six tasks used to study rhetorical task and topic familiarity. Each student wrote only one essay. Three participants' data were discarded: one student produced a list of ideas in bullet points instead of an essay, and two students obviously did not treat the essay task seriously and produced only 3-4 sentences showing no attempted efforts. The students were recruited from intact classes from a major public university located in Southeast China. Among the 61-64 students writing on each task, approximately 30 were English-major freshmen and sophomores with approximately 10 freshmen and 20 sophomores, and approximately 30 were non-English-major freshmen and sophomores with approximately 10 freshmen and 20 sophomores. At the time of the data collection, all the students were taking English classes from the university, with the classes separated by major–English or non-English and by grade level–freshmen or sophomores. The sampling technique aimed to have writers with a good range of L2 proficiency levels for each task and to obtain samples of writers with equivalent language proficiency ranges across the six tasks. The six essay tasks were randomly distributed within each of the 15 participating English classes, so that within each class, an approximately equal number of essays were collected for the six tasks. Such random distribution was to further ensure that the groups for the six writing tasks were equal, not influenced by particulars of intact classes. The students completed the writing tasks in class.

Table 3.1 below displays the summary of demographic information of the participants. The participants all filled out a brief demographic information questionnaire before the writing

Table 3.1
Participant Characteristics

	Characteristic	*N*
Total		372
Age*	Mean (years): 20	
	Range: 17-24	
Gender*	Female	228
	Male	142
Status & Major	English Freshmen	67
	English Sophomore	118
	Non-English Freshmen	61
	Non-English Sophomore	126
Non-English Major Field of study*	Business	28
	Computer Science	30
	Engineering	56
	Humanities	2
	Law	9
	Mathematics	11
	Natural Sciences	13
	Social Sciences	31
# of years studying English	Mean: 8 years 10.99 months	
	Range: 4 years-15 years	
study- or travel-abroad experience in English-speaking countries	Yes	0
	No	372

*Age information missing for 6 participants.
Gender information missing for 2 participants.
# of years studying English missing for 10 participants.
Non-English major missing for 7 participants.

task, covering information about the students' name, gender, age, academic status, academic major, number of years of learning English, and length of stay in English-speaking countries (if any). See Appendix B for the demographic information questionnaire. The students could choose to complete either the English version or the Chinese version of the questionnaire.

3.2 Research Materials

3.2.1 Writing tasks

A total of six writing tasks were used in the study, being all on the same subject matter and varying along two dimensions: rhetorical task and topic familiarity. Four different prompts were used for the four rhetorical tasks (i.e., narration, exposition, expo-argumentation, and argumentation) with topic familiarity controlled at the medium-impersonal familiar level except for the narrative task. Three different prompts were used for the three levels of topic familiarity (i.e., higher–personal familiar topics, medium–impersonal familiar topics, and lower–impersonal less familiar topics), with rhetorical task controlled at the expo-argumentative task level. One writing prompt which was an expo-argumentative task with medium topic familiarity was shared between the two prompt sets for the two cognitive complexity dimensions. Control over the other dimension in a study of the effect of one of the dimensions is much recommended, since an interaction effect has been observed for the two dimensions (Hamp-Lyons & Mathias, 1994). Table 3.2 lists the six writing prompts that were used, all sharing the common subject matter of the use of computers and the Internet. Control over subject matter was considered necessary since topic/ subject matter has been found to have an effect on writing performance in scores and in language production features, as discussed in Chapter 2. What is construed in each of the six prompts is different, although the same subject matter is used. A fully crossed design was not feasible, due to the number of levels for the two dimensions: 4×3.

Table 3.2
Writing Prompts Used in the Study

Familiarity/ Rhetorical Task	Personal Familiar	Impersonal Familiar	Impersonal Less Familiar
Narrative	Describe **one** of your experiences in which you used computers and/or the Internet for completing a course assignment or project or for studying for a school subject matter.	--	--
Expository	--	What are some ways that university students in this country use computers and the Internet?	--
Expo-argumentative	What do you think are the benefits and possible problems that computers and the Internet bring to **you** as a university student?	What do you think are the benefits and possible problems that computers and the Internet bring to university students in this country?	What do you think are the benefits and possible problems that computers and the Internet bring to people in underdeveloped areas of the world where there is limited access to computers and the Internet?
Argumentative	--	Computers and the Internet have improved the efficiency and quality of learning for university students in this country. Do you agree or disagree with the statement? Support your position with reasons.	--

The writing task sheet that was administered to the participants can be found in Appendix C. The writing tasks and the directions for the essay task were given in both English and Chinese. The essay writing was timed, and the students were given 30 minutes to write their essays. 30 minutes is also the time limit for the writing sections in College English Test and in Test for English Major in China, as well as the independent writing section in Test of English as a Foreign Language (TOEFL). In the directions for the writing tasks, the students were also instructed to write a minimum of 150 words and were informed of the broad areas on the basis of which their essays would be rated: idea development and support in relation to the prompt and the task, organization and flow of ideas, and language use (in syntax, lexis, and etc.).

The participants hand wrote the essays. After the hand-written essays were collected, the essays were typed up into Microsoft Word files onto computers, by several hired students at the participating university and by the researcher. The researcher then checked each single typed-up essay against the original hand-written essay for accuracy, and any discrepancies were corrected.

3.2.2 Post-writing questionnaire

A short post-writing questionnaire was administered right after the writing task was completed. The questionnaire mainly asked questions about the writers' pre-writing planning for the task completed, level of interest in the writing topic, level of familiarity with the rhetorical task, frequency of use of computers and the Internet, perception of difficulty levels of the writing task completed and the other writing tasks in the same cognitive complexity category. See Appendix D for the questionnaire. The students could choose to complete either the English version or the Chinese version of the questionnaire.

3.2.3 Measure of general English proficiency

Since the study also looks into whether general English proficiency plays a role in the effects of the cognitive complexity dimensions on writing quality and on language production features of writing, a measure of the participants' general English proficiency was needed. For this purpose, a cloze test that was first validated by Brown (1980) was used; it is a 50-item test that requires 25 minutes. A cloze test was chosen because it is often found to be an adequate indicator of general language proficiency (Brown, 2002; Hinofotis, 1980; Oller & Conrad, 1971). Other longer tests of general English proficiency could not be given in the study due to time constraints in the classrooms. The cloze test used, with translation of the directions in Chinese, and the answer keys are in Appendix E. The participants were given 30 minutes, since the cloze test format was not familiar to the participants. The researcher, together with a visiting scholar from China, scored the cloze tests using acceptable answer scoring, with the score for each test checked twice when scoring. For the whole data set, the mean score of the cloze test was 27.33, the median was 27, and the range was 6-47. The cloze test was to further confirm that the task groups were equal in terms of their general L2 proficiency level. Table 3.3 below reports the descriptive statistics for the cloze test results for each of the writing task groups. The mean scores for the groups were almost identical, and there were no statistical differences in the means, confirming that the L2 proficiency levels of the task groups were equal.

Table 3.3

Means and Standard Deviations for Participants' Cloze Test Scores by Writing Task Group

Group	*n*	*M*	*SD*
Narrative	61	28.02	8.21
Expository	62	27.05	8.58
Expo-Argu/ Impersonal Familiar	61	26.66	8.51
Argumentative	63	26.65	9.76
Personal familiar	63	27.75	7.78
Impersonal less familiar	62	27.87	8.29

3.3 Recruitment and Consent Procedures

The English teachers of the participating classes recruited the participants and conducted the consent procedures, following the instructions from the researcher. The researcher specified the actual verbal announcements for the recruitment and the consent procedures. The verbal announcements were written in both English and Chinese and can be found in Appendix F. The instructors were asked to announce the recruitment and the consent procedures using the Chinese version, and the Chinese version of the IRB-approved consent form was used. The students in the participating classes, after reading the research consent form, either chose to participate in the study and completed the research procedures or chose not to participate; for those who chose not to participate, they could still complete the instruments for the study but their data were not collected, or their instructors assigned other course-related tasks for them to complete during the research sessions.

3.4 Data Collection Procedures

Data collection took place in two class sessions of intact classes. The English teachers of the participating classes administered the research materials, following the written instructions provided by the researcher. The written instructions given to the instructors about the research procedures were in both English and Chinese and can be found in Appendix G. In the first class session, the participants completed the consent procedures (5 minutes) and then the cloze test (30 minutes). In the second class session, the participants filled out the brief demographic information questionnaire (3 minutes), completed the writing task (30 minutes), and then filled out the post-writing questionnaire (5 minutes). The class sessions are 40 minutes each in the participating university.

3.5 Essay Rating

3.5.1 Writing quality rating of essays

All the essays, in their original hand-written format, were scored by trained raters, using the TOEFL iBT Test Independent Writing Rubrics (ETS, 2012). The original rubrics have six scale points, from zero to five; in this study, a half point was added between score points, and a description of "**Half-point ratings** (e.g., 2.5) are given when an essay's quality falls in between the descriptors for two adjacent whole points" was added to the original rubrics. The primary reason for adding half-point ratings was that the English language proficiency of the participants was not widely dispersed, with a proficiency range from low intermediate to low advanced, and with the addition of half-point ratings, more score points could be generated to reflect finer-grained writing quality differences and to assist with the detection of significant findings if there is any. Secondly, based on the descriptors of the rating rubrics, half-point ratings could be naturally assigned to the essays which did not completely fall into the descriptors for a higher whole-point or the lower whole-point. Both the researcher and the raters found it reasonable and easy to use the half-point ratings added. Appendix H shows the rubrics used, with the half-point rating description and slight formatting of the original rubrics. The researcher selected from the collected data sample essays representative of major score points based on the rubrics, in consultation with several graduate students and scholars specializing in language assessment, and used those sample essays for the training and norming sessions she conducted.

Raters for the study were recruited from the rater pool for the Georgia State Test of English Proficiency (GSTEP) and were paid with an hourly rate of $15. There were a total of five raters: four raters were PhD students in the Department of Applied Linguistics and ESL and one rater was a senior lecturer in the Intensive English Program at Georgia State University. All the raters had previous experience in rating ESL essays, and all had ESL/EFL teaching experience. All the raters went through the training and norming sessions for the study before the actual scoring. Essays on the six tasks were all mixed together when they were rated, so that the raters might be less influenced by particular task characteristics and task difficulty. Each essay was rated by two raters, and when there was a discrepancy of more than one point between the two raters for an essay, a third rater also rated the essay. For each essay, the average of the two ratings whose difference was one point or less than one point was taken as the final score for that essay. A maximum total of 21 final score points can result from the scoring procedure: 0, 0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75, 2, 2.25, 2.5, 2.75, 3, 3.25, 3.5, 3.75, 4, 4.25, 4.5, 4.75 and 5. The actual score range was 1.25 to 5 for this study, with a median of 3. The inter-rater reliability for the whole data set, as measured by Cronbach's α, was .84. The inter-rater reliability for each task ranged from .80 to .88, with the lowest reliability (.80) found for the narrative task and the personal-familiar task and higher reliabilities (.84 - .88) found for the other four tasks.

3.5.2 Task fulfillment rating of essays

Task fulfillment rating was done for each of the six tasks used in this study, mainly to see whether the writers approached the task as asked, e.g., whether the writers wrote a narrative essay for the narrative task and whether they wrote an expository essay with exposition defined in this study. The researcher read through all the essays collected on each task and created the task fulfillment rating rubrics based on the actual essays produced and the approaches taken. The rating rubric for each task is provided in Appendix I, and the rating scales for the tasks are all categorical. Listed in Table 3.4 below are the codes and descriptors for the fulfilled tasks for each task; see the appendix for the complete rating scales.

Table 3.4
Codes and Descriptions for Fulfilled Tasks for the Task Fulfillment Rating

code	**description**
Narrative	
1	The writer mainly described **one** of his/her experiences in which he/she used computers and/or the Internet for completing a course assignment or project or for studying for a school subject matter.
Expository	
1	The writer mainly described and made generalizations of ways that university students use computers and the Internet, although occasional judgment of the uses as being positive and/or negative might be involved.

Expo-argumentative/ Impersonal-familiar	
1	The writer primarily approached the task collectively (from the perspectives of the 1st person plural – we, us, our, ours, and/or "university/college students" in general or by subgroup and/or the 3rd person plural – they, them, their, theirs) and discussed the benefits and problems that computers and the Internet brought to them and/or university/college students.
Argumentative	
1	The writer clearly stated his/her position on the debatable statement by choosing a side and wrote his/her support for the position.
2	The writer stated his/her position on the debatable statement by providing conditions for the truth value of the statement and wrote his/her support for the position.
Personal-familiar	
1	The writer primarily approached the task personally (from the perspectives of the 1st person singular – I, me, my, mine) and discussed the benefits and problems that computers and the Internet brought to him/her.
Impersonal-less familiar	
3	The writer considered and adequately addressed the issue from the perspectives of the people in the underdeveloped areas. There was clear and adequate indication that the writer's treatment of the issue was primarily context-specific, with **adequate and explicit** verbal contextualizations for **both** the benefits and problems of computers and the Internet for the context specified.

The researcher, together with another PhD student, rated the essays with the rubrics. The other rater is also an experienced ESL teacher and an experienced writer. For all the tasks except for the impersonal-less familiar task, we first independently rated each essay and assigned rating categories, and then we resolved differences through discussions. The agreement for the initial independent ratings for the five tasks was in the range of 90%-99%. For the impersonal-less familiar task, the two raters had substantial disagreement in ratings with an originally developed task fulfillment rating rubric for that task. The rubric finally developed and used for this task was based on further examinations of the original rating rubric and discussion between the raters, and it was much easier to use. However, since the final rating rubric for this task was in the sense of the degree to which the writers were considering the issues in relation to the less familiar context and were thus truly addressing a less familiar topic and the raters had their own judgments and had more disagreement than they did for the other tasks, it was decided that there did not need to be complete agreement between the two raters for each rating for this task. Therefore, after each rater independently rated the essays for this task, no discussion was pursued when there were discrepancies in the ratings. 86% of the independent ratings were in total agreement for this task, and the rest of the 14% were all one category apart, showing no large disagreement.

Since whether the writers approached the tasks as asked, thus engaging in the primary cognitive processes the tasks were designed to invite, was related to the research questions of the current study, the proportions of the fulfilled tasks based on the task fulfillment rating results are reported and summarized here. Table 3.5 below displays the proportion of fulfilled tasks for each of the six tasks. The results of the task fulfillment ratings were primarily to inform whether separate data analyses were necessary for the on-task sample only for a given task, since the on-task essays can more truly reflect the influences of the cognitive demands of the tasks. As Table 3.5 shows, the writers fulfilled all the four rhetorical tasks well, with 89%-97% of the writers completing the tasks as asked. Since the writers were able to fulfill all the four rhetorical tasks well, no separate analyses for the on-task samples only were deemed necessary for these tasks. For the topic familiarity dimension, the writers fulfilled the impersonal-familiar expo-argumentative task well, which was a shared task with the rhetorical task dimension. However, only 33% of the writers for the personal-familiar topic wrote personal essays, with most of the rest producing impersonal essays, and only 56% of the writers for the impersonal-less familiar topic produced essays that showed that they were evidently and adequately addressing the less familiar context, with the rest of them showing limited or almost no evidence in doing so and making the less familiar topic a more familiar one. These 56% of the essays on the impersonal-less familiar topic were based on ratings that both of the raters agreed to be a 3, not ratings that one rater assigned 3 and the other assigned 2, thus only including the essays that both raters thought were truly addressing a less familiar topic. Based on these results, separate data analyses for the on-task samples only were conducted for the personal-familiar and the impersonal-less familiar tasks in the familiarity dimension, when the sample sizes were sufficient.

Table 3.5
Proportion of Fulfilled Tasks Based on the Task Fulfillment Rating

		Code for Fulfilled Tasks*	Proportion of Fulfilled Tasks
Rhetorical task	Narrative	1	57/61 (93%)
	Expository	1	55/62 (89%)
	Expo-argu	1	57/61 (93%)
	Argumentative	1; 2	61/63 (97%)
Topic familiarity	Personal-Familiar	1	21/63 (33%)
	Impersonal-Familiar	1	57/61 (93%)
	Impersonal-Less familiar	3	35/62 (56%)

* The descriptors of the codes can be found in Table 3.4 above.

3.6 CAF Measures

The language performance features the study investigated included complexity, accuracy, and fluency of language production in the ESL essays. Linguistic complexity comprises of lexical complexity and syntactic complexity. Specific measures used for each area, as well as the tools to obtain the measures, are outlined below.

3.6.1 Accuracy measure

For linguistic accuracy, total number of grammar and usage errors per 100 words as measured by e-rater for each essay was used as the measure. As laid out in the literature review section, a normalized total number of errors (or errors in a specific area, e.g., grammatical errors) is used as an accuracy measure in both the TBLT and the writing literature (e.g., Arnaud, 1992, Linnarud, 1986; Sangarun, 2005). In this study, total numbers of grammar errors and usage errors for each essay were automated and obtained through the e-rater engine (version13.1) of

Educational Testing Service (ETS), after a research use request was approved by the ETS. The ETS Innovations in the Development and Evaluation of Automated Scoring Group processed the typed-up essays in their original form with the e-rater engine. The e-rater engine is “an ETS capability that identifies features related to writing proficiency in student essays so they can be used for scoring and feedback” (ETS, 2014), and it is capable of detecting errors in the areas of grammar, usage, and mechanics. Errors are automatically detected when frequencies of certain bigrams or trigrams in student essays are greatly lower than those identified through in a large corpus of texts on which the engine is trained (Leacock & Chodorow, 2003). The error types that are detectable through e-rater cover a good range. For example, grammar errors include ill-formed verbs, wrong or missing words, run-on sentences, subject-verb agreement, possessive errors and so on, and usage errors include article errors, incorrect word forms, confused words and so on (See Quinlan, Higgins, & Wolff, 2009 for a complete list for e-rater 2.0 capacity). The e-rater engine which processed the essays is version 13.1, and the types of errors that it is able to detect have further expanded, with nine types of grammar errors and 10 types of usage errors (Personal communication, ETS Innovations in the Development and Evaluation of Automated Scoring Group). The automation of error counts that the e-rater engine provided allowed for a good survey of errors made in the essays in an efficient and reliable manner. After the total number of grammar errors and the total number of usage errors in each essay were obtained from e-rater, these totals were added, then divided by the word count of each essay and finally multiplied by 100 to generate the accuracy indices–total number of grammar and usage errors per 100 words.

3.6.2 Fluency measure

For fluency of language production, the total number of words in each essay was used as the measure. Since the essay writing is only limited to 30 minutes, the total number of words generated in the given time is a good measure for fluency. Text length is a main fluency measure that has been used in L2 writing studies (Wolfe-Quintero, et al., 1998), and is more valid than the measures of total number of T-units or clauses (Polio, 2001). Essay length was obtained through word count in Microsoft Word.

3.6.3 Lexical complexity measures

Lexical complexity was captured through the sub-constructs of lexical diversity, lexical sophistication, and lexical density. To measure lexical diversity, the vocd D measure (Malvern, et al., 2004) from the Computerized Language Analysis (CLAN) programs of the CHILDES project (MacWhinney, 2000) was used. Unlike many lexical diversity measures such as type-token ratio, vocd D measure is less sensitive to text length since it is based on 100 times of random sampling for each of the 35-word up to 50-word text portions in a text, and it is reported to be reliable for text lengths with ranges of 100-400, 200-500, 250-666, and 400-1000 (McCarthy & Jarvis, 2007). This measure has been widely used in L1 studies and is finding its popularity in L2 studies. The vocd D measure suits the essay data for this study well, since all but two essays collected fall within the range of 100-400 words.

For lexical sophistication, the proportion of sophisticated word types in each essay was used as the measure. The concept of sophisticated words is based on Laufer and Nation (1995) in which words beyond the most frequent 2,000 words in English are classified as more advanced, sophisticated and lower frequency words. The sophisticated words include most academic vocabulary, domain-specific words, as well as other less frequently used words. In the current study, the most frequent 2,000 lemmas selected for use as the basic words are based on the second release of the American National Corpus (ANC; Reppen, Ide, & Suderman, 2005). Currently, the open ANC contains 15 million words of contemporary American English from written and spoken texts of all genres produced since 1990, and the second release of ANC contains 22,000,000 words from the full corpus which are annotated for lemma, part of speech and so on (ANC, 2014). To obtain the indices for the proportion of sophisticated word types in each essay, the Lexical Complexity Analyzer (Lu, 2012) was used; the software automates the measure and provides counts of total word types and sophisticated word types based on the ANC most frequent 2,000 lemmas. Following Laufer and Nation (1995), proper nouns in each essay that were not in the most frequent 2,000 lemmas were deleted from the sophisticated word types. Proper nouns in each essay were automatically identified through the RANGE program (Heatley, Nation, & Coxhead, 2002), after the essays were processed through the program in batches; with the default option, proper nouns appear in the “Types Not Found In Any List” for each essay. A list of the ANC most frequent 2,000 lemmas was obtained from Lu, the author of the Lexical Complexity Analyzer. In addition to the proper nouns, for each essay, prompt words that were not in the most frequent 2,000 lemmas were also deleted from the sophisticated word types. There were a total of four prompt words that were not in the most frequent 2,000 lemmas: *assignment* in the narrative prompt, *efficiency* and *disagree* in the argumentative prompt, and *underdeveloped* in the lower topic familiarity prompt. For each essay, the number of word types for the relevant proper nouns and prompt words was subtracted from the number of sophisticated word types in each essay, and the resulting number was then divided by the total number of word types in the essay to generate the final index of the proportion of sophisticated word types.

Finally, the lexical density measure–the number of lexical words out of the total number of words in a text, was used, and indices for the measure were obtained from the Lexical Complexity Analyzer (Lu, 2012), which automates the measure. Lexical words are contrasted with functional or grammatical words such as articles, prepositions and pronouns. Although such a contrast is largely accepted, the previous literature defines and counts lexical words with certain variability. In Lu (2012), lexical words include "nouns, adjectives, verbs (excluding modal verbs, auxiliary verbs, "be," and "have"), and adverbs with an adjectival base …" (p. 192).

For all the lexical complexity measures used in the current study, based on the procedure in Laufer and Nation (1995), for each essay, all spelling errors were corrected before indices were obtained from the computer programs.

3.6.4 Syntactic complexity measures

Eight different measures were used for syntactic complexity (SC), representing different dimensions of the multi-dimensional construct (Norris & Ortega, 2009). The eight measures used were: two global SC measures–mean length of sentence (MLS) and mean length of T-unit (MLTU), one clausal coordination measure–T units per sentence (TU/S), one measure tapping into overall clause complexity–mean length of clause (MLC), two subordination measures–finite dependent clauses per T-unit (DC/TU) and nonfinite elements per clause (NFE/C), one phrasal coordination measure–coordinate phrase per verb phrase (CP/VP), and one noun-phrase complexity measure–complex noun phrases per verb phrase (CNP/VP). The definitions of the eight measures and the sub-constructs they represent are summarized in Table 3.6 adapted from Yang, Lu, and Weigle (under revision).

Table 3.6

Syntactic Complexity Measures Used

Sub-construct	Measure	Definition
Overall sentence complexity	Mean length of sentence (MLS)	Number of words divided by number of sentences
Clausal coordination	T-units per sentence (TU/S)	Number of T-units divided by number of sentences
Overall T-unit complexity	Mean length of T-unit (MLTU)	Number of words divided by number of T-units
Clausal subordination	Dependent clauses per T-unit (DC/TU)	Number of dependent clauses divided by number of T-units
Overall clause complexity	Mean length of clause (MLC)	Number of words divided by number of clauses
Phrasal coordination	Coordinate phrases per clause (CP/VP)	Number of coordinate phrases divided by number of verb phrases
Noun phrase complexity	Complex NPs per verb phrase (CNP/VP)	Number of complex NPs divided by number of verb phrases
Non-finite elements/subordination	Non-finite elements per clause (NFE/C)	Number of non-finite elements divided by number of clauses

Figure 1 below, adapted from Yang, Lu, and Weigle (under revision), graphically shows the hierarchical relationships among the SC sub-constructs and their measures and how they represent SC as a multi-dimensional construct. The current study, following the definitions given in grammar theories (Cristofaro, 2003; Givon, 2008; Halliday & Matthiessen, 2004; Langacker, 2008), regards both finite and nonfinite dependent structures as subordination and thus examines both DC/TU and NFE/C. Although previous writing studies have frequently examined the amount of finite subordination, non-finite subordination has not received due attention. Further, following the existing writing literature, clause in this study refers to only finite clauses (see Hunt, 1965; Lu, 2011; Norris & Ortega, 2009; Polio, 1997), and nonfinite structures are thus referred to as non-finite elements.

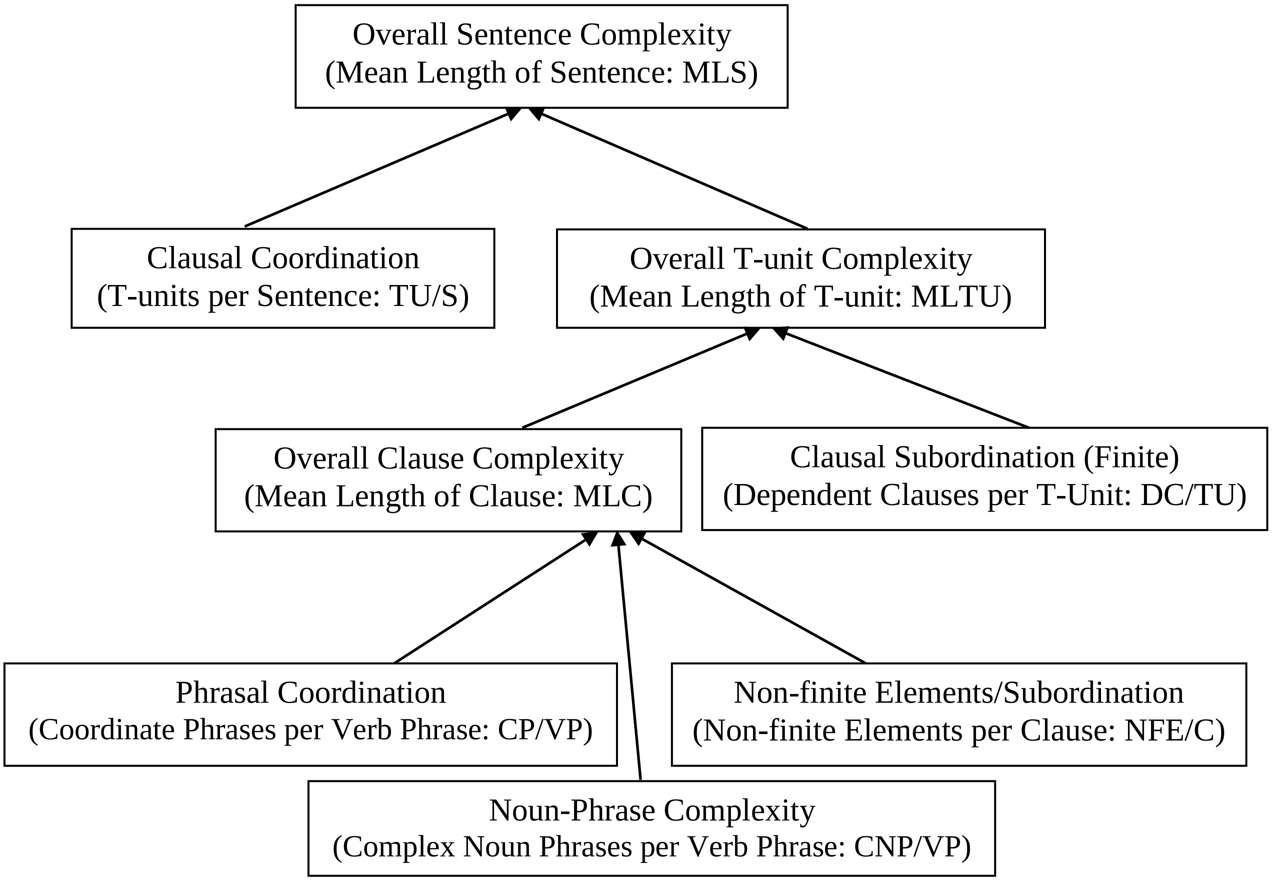
Overall Sentence Complexity
(Mean Length of Sentence: MLS)
Clausal Coordination
(T-units per Sentence: TU/S)
Overall T-unit Complexity
(Mean Length of T-unit: MLTU)
Overall Clause Complexity
(Mean Length of Clause: MLC)
Clausal Subordination (Finite)
(Dependent Clauses per T-Unit: DC/TU)
Phrasal Coordination
(Coordinate Phrases per Verb Phrase: CP/VP)
Non-finite Elements/Subordination
(Non-finite Elements per Clause: NFE/C)
Noun-Phrase Complexity
(Complex Noun Phrases per Verb Phrase: CNP/VP)

Figure 3.1. A Multi-dimensional Representation of Syntactic Complexity

Indices for the eight measures of SC for each essay were automated with a computational tool–L2 Syntactic Complexity Analyzer (L2SCA) (Lu, 2010), with some necessary minor adaptations as described below. Explicit definitions of the linguistic units relevant to this study and automated in L2SCA –sentence, T-unit, clause, dependent clause, coordinate phrase, and verb phrase–can be found in Lu (2010, 2011). The original version of L2SCA provides frequency counts of the above linguistic units and other linguistic units and generates 14 different SC indices for a given text. The original version of L2SCA generated indices for MLS, MLTU, T/S, DC/TU, and MLC for each essay in this study. CP/VP was calculated by dividing the frequency counts of coordinate phrases (CP) by the frequency counts of verb phrases (VP). Complex noun phrases (CNP) are defined in this study as noun phrases that contain one or more of the following: pre-modifying adjectives, post-modifying prepositional phrases, and post-modifying appositives (see, e.g., Biber, Gray, & Poonpon, 2011). For this study and for Yang, Lu, and Weigle (under revision), Lu, the author of L2SCA, used the pattern for identifying complex nominals in the original L2SCA to modify accordingly to match this definition of complex noun phrases in order to automate frequency counts of CNP. CNP counts were then divided by VP counts to generate CNP/VP for each essay. Finally, L2SCA calculates verb phrases per clause (VP/C) but not non-finite elements per clause. The number of non-finite elements per clause was computed by subtracting 1 from VP/C. This was done because in L2SCA, each clause contains one finite VP, and the other VPs are therefore non-finite. As for the reliability of L2SCA, Lu (2010) reports that the tool is highly reliable in the production-unit frequency counts and the SC indices generated for college-level ESL writing at the intermediate and high proficiency levels, based on an analysis of and comparison with human coding of sample essays produced by Chinese college-level EFL writers.

3.7 Data Analysis

To answer research questions 1 and 4 regarding the effect of the cognitive complexity of tasks on writing scores, one-way ANOVA (Analysis of Variance) tests were separately conducted for the two independent variables: rhetorical task and topic familiarity. Then to see whether general L2 proficiency would make a difference in the results, two-way ANOVA tests were separately conducted for each of the cognitive complexity dimensions, with general L2 proficiency based on cloze test scores and divided into lower- and higher-proficiency groups using the median of the cloze test scores for the whole sample collected (*n* = 373) as the cut point.

To address research questions 2 and 5 about the effects of the cognitive complexity of tasks on CAF of language production, MANOVA (Multiple Analysis of Variance) tests were first separately conducted for rhetorical task and topic familiarity with their multiple dependent variables to see whether these variables had an effect on the CAF features at the omnibus level. After this, when significant results were revealed through the MANOVA analyses, separate univariate analyses were conducted to see the effect of each of the cognitive complexity dimensions on each of the measures for accuracy, fluency, lexical complexity, and syntactic complexity, with the α level adjusted based on Holm procedure due to the multiple univariate tests. Further to see whether general L2 proficiency would make a difference in the results, two-way MANOVA tests were separately conducted for the two cognitive complexity factors, with general L2 proficiency groups identified as described above. Follow-up univariate analyses were further pursued.

For research questions 3 and 6 examining the predictive power of CAF on writing scores, first, bivariate correlations using Pearson product-moment correlation were run for each task to see the strength of the relationship between each of the CAF variable and writing scores, as well as that among the CAF variables. Then separate multiple regression analyses using all-possible-subsets regression were conducted for each task, with writing scores as the dependent variable and selected CAF features as the predictor variables. The best regression models and the predictive power for the four levels of rhetorical task were compared against each other, and the same were done for the three levels of topic familiarity.

All-possible-subsets regression, in contrast to the often-used step methods (forward, backward, and forward stepwise), makes possible an exhaustive analysis of all subsets (often combinations) of predictor variables and their predictive power, including 0, 1, 2, 3, 4, 5 and so on predictors. For *X* number of predictors in the full model, there are a total of 2^x subsets of possible regression models. Often reported along with all-possible-subsets regression analysis are information criteria that can help researchers to make decisions as to which combinations of predictor variables are the most parsimonious ones that can predict the dependent variable well, as well as the full model. The information criteria can be calculated automatically by current versions of popular statistical analysis programs (e.g., SPSS, SAS, MINITAB) or with the assistance of computer programs such as Excel. Instead of presenting only one regression model as the step methods do, the all-possible-subsets regression method is able to provide the researcher with several regression models that can predict the dependent variable well, often as well as what the step methods may produce (Huberty, 1989; Kutner, Neter, Nachtsheim, & Li, 2005; Stevens, 2009). It is then the researcher's further decision to choose the best regression model based on other diagnostic information, his or her expert knowledge on the subject matter, and the purpose of the analysis (Kutner, Neter, Nachtsheim, & Li, 2005).

The information criteria that can help the researcher to make the decision about the "best" regression model mainly include R^2, adjusted R^2, MSE (mean squared error), Mallows' Cp, AIC (Akaike's Information Criterion), SBC (Schwarz' Bayesian Criterion), and PRESS (prediction sum of squares) (See Kutner, Neter, Nachtsheim, & Li, 2005 for a review). Since AIC and Mallows' Cp are the primary ones used in the current study, an explanation of them follows. The calculation of AIC includes the SSE (sum of squares for the error) for a specific subset of predictors, the number of predictors, and the sample size. With the SSE or the number of predictors increasing, the AIC is likely to increase. The smaller the AIC is, the better a model is. The idea is that the best models are the ones that have an SSE as small as the one for the full model, with a smaller number of predictors. AIC thus penalizes having more predictor variables than necessary. AICC, or AIC corrected (Hurvich & Tsai, 1989), is recommended for smaller sample sizes and is thus adopted in this study. Mallows' Cp (Mallows, 1973) estimates the bias of each subset of predictors, generally by assuming the full model as the unbiased one. The calculation of Mallows' Cp includes the SSE for a specific subset of predictors, the MSE of the full model, the number of predictors in the subset, and the sample size. Larger SSE for a subset of predictors will increase Mallows' Cp. It is recommended that when Mallows' Cp is approximately equal to k +1 (k = number of predictors), a subset model shows little or no bias, when Cp is much larger than k +1, there is great bias in the model, and when Cp is smaller than k +1, the model shows no bias (Kutner, Neter, Nachtsheim, & Li, 2005). In general, the smaller the Cp is, the better a model is (Stevens, 2009).

Since each task in this study only had approximately 60 subjects, the number of predictors that were entered in the regression analyses was limited to five. Statisticians have suggested having approximately 15 or more subjects per predictor for social science studies for a reliable regression equation (see Stevens, 2009) or a minimum of 6-10 in a general sense (Kutner, Neter, Nachtsheim, & Li, 2005). The five predictors were selected based on construct representation, to make sure that all the CAF feature areas of accuracy, fluency, lexical complexity and syntactic complexity were represented. Further, the results of the bivariate correlations between CAF variables and writing scores were used to inform the selection of the five predictors, as well as related previous literature on the relationships. In this study, all possible subsets regression analyses were conducted with an Excel program made by Oshima

(2013); since the final predictors that were included in the analyses were five, the program was a good solution for the analyses, after some minor expansion of predictor variables. Although SPSS version 19 and above are capable of running all-possible-subsets regression through its Automatic Linear Modeling function, this function has its limitations in not reporting the informative indices of Mallows' Cp and adequate R^2 which are also useful for model selection.

Before Oshima (2013) was run for each task, first for each combination of predictor variables, including one predictor, linear regression analysis was conducted through SPSS version 18, to obtain the SSE (sum of squared errors) and the R^2 for the predictor combination. The indices for each predictor combination were then inserted into the corresponding cells in the Excel program row by row, along with the number of predictors for each combination. In addition, the total sample size, the total number of predictors, and the MSE (mean squared error) and the R^2 for the full model (i.e., the model with all five predictors) were inserted into the appropriate cells. After all these indices were provided, the program automatically calculated, with its pre-inserted formulas, the adjusted R^2, the MSE (mean squared error), Mallow's Cp, AIC, and AICC for each predictor combination, as well as an indication of whether the R^2 for each predictor combination was greater or smaller than the adequate R^2 based on the full model. The adequate R^2 value was also automatically calculated by the program, and any value below it was statistically smaller than the R^2 of the full model. The original Excel program has a maximum of four predictor variables, and the researcher expanded the program to a total of five predictor variables.

Before each of the above statistical analyses was conducted, assumptions for the analyses were first checked, and they were all found to be met unless specified in the Results chapters. When assumptions were not met, other appropriate statistical analyses were used.

Since outliers can have a great effect on the statistical findings and the task comparison results, all the analysis of variance tests for research questions 1, 2, 4, and 5 were also conducted by removing outliers from the analyses, with outliers defined as the ones lying 3 standard deviations above or below the mean ($z > 3$ or $z < -3$) for each variable for each task. For example, if a writer who wrote on the narrative task produced an essay with a length that is 3 standard deviations above or below the mean of essay length for the narrative task, the writer would be treated as an outlier for the fluency variable. The outlier analyses revealed that most of the variables in the study had very few outliers for the whole data set, ranging from 2 to 4. Writing scores, cloze test scores, lexical diversity measure, and lexical density measure did not have outliers. The one measure that had more outliers is the SC measure of T units per sentence (TU/S); there were a total of seven outliers. The primary reason for this was that six out of the seven outliers produced run-on sentences. An inspection of the other outliers' data and essays showed that the outliers were exceptional cases on certain variables, may not be representative of the sample they belonged to for those variables, and could affect the statistical testing results. For example, the narrative, expository, and argumentative tasks each had an outlier for the fluency measure, while the expo-argumentative task did not have an outlier for this measure; the outliers could potentially affect the results when these tasks were compared on the fluency measure.

When the analyses were conducted without the outliers, the data were removed casewise rather than listwise, since in general there were no good reasons to remove all the data for an essay if it was an outlier on one language production variable such as lexical sophistication. However, due to the inter-connectedness of the syntactic complexity variables (See Figure 1), one SC measure is likely to have an effect on the magnitudes of some other SC measures. For example, a subject who overtly uses nonfinite subordination not only is likely to be outlier on nonfinite elements per clause but also could have longer mean length of clause, mean length of T-unit, and mean length of sentence due to the much greater use of nonfinite subordination, making the data for these other variables not representative of typical cases. It was thus decided to remove all the SC indices for a subject who was identified as an outlier on one of the SC measures.

The outliers for the multiple regression analyses for research questions 3 and 6 were identified differently, since there are other recommended procedures for such a purpose. First, the outliers based on z scores (z > 3 or z < -3) for each variable were still removed. Then for each task's five predictor variables, indices for Cook's D and Standardized DfBeta (DFBETAS), two recommended leverage measures (e.g., Yang, 2013) showing the influences of each specific case on the regression coefficients, were obtained from SPSS version 18. A participant was identified as an outlier for the regression analysis, when its Cook's D value is greater than 4/*n* (*n* = sample size), which is the standard practice for the cut-off point for Cook's D, and when its DFBETAS value is greater than 1, which is the recommended cut-off point for small-to-medium sample sizes (Kutner, Neter, Nachtsheim, & Li, 2005). The outliers identified through Cook's D and DFBETAS values were also removed for the regression analyses without the outliers.

CHAPTER 4: RESULTS AND DISCUSSION FOR THE RHETORICAL TASK DIMENSION

This chapter presents the results for the rhetorical task cognitive complexity dimension, answering research questions 1, 2, and 3 presented in Chapter 1. Specifically, it presents the results about the effect of rhetorical task on L2 writing scores for college-level EFL students, the effects of rhetorical task on accuracy, fluency, and linguistic complexity (i.e., lexical complexity and syntactic complexity) of L2 writing production, and the predictive power of the CAF features on L2 writing scores for the different rhetorical tasks. The four rhetorical tasks examined were narrative, expository, expo-argumentative, and argumentative tasks. A discussion of the results follows, in relation to the hypotheses proposed and previous related studies. Implications of the findings for L2 teaching and assessment and theorizing of task cognitive complexity are laid out in the concluding chapter.

4.1 Results

4.1.1 Effect of rhetorical task on L2 writing scores

Table 4.1 below displays the descriptive statistics for the L2 writing scores on the four rhetorical tasks. A one-way ANOVA test revealed that there was no statistical difference among the group means, $F(3, 243) = .36$, $p = .78$, $\eta^2 = 0.004$, with an achieved power of 0.85 for the statistical analysis. According to Cohen (1977) and Stevens (2009), eta squared (η^2) and partial eta squared ($\eta^2_{partial}$) of the value of .01 is considered as small, .06 as medium, and .14 as large. The effect size observed for this comparison is negligible, particularly in view of the adequate power achieved, showing that even with a larger sample size, a difference is unlikely to be observed. The participants performed equally well on the four different rhetorical tasks varying in reasoning demands.

Table 4.1

Means and Standard Deviations for L2 Writing Scores by Rhetorical Task

Group	*n*	*M*	*SD*
Narrative	61	2.94	0.76
Expository	62	3.00	0.77
Expo-Argu	61	2.98	0.74
Argumentative	63	3.08	0.74

Table 4.2 below presents the descriptive statistics for the L2 writing scores on the four rhetorical tasks, divided based on the two L2 proficiency levels determined by the cloze test. A two-way ANOVA test, with task and L2 proficiency level as the independent variables and L2 writing scores as the dependent variable, was conducted. There was no main effect from the rhetorical task, $F(3, 239) = 1.43$, $p = .23$, $\eta^2_{partial} = .02$, with an achieved power of .38. A significant main effect was found for the L2 proficiency level, $F(1, 239) = 109.37$, $p < .01$, $\eta^2_{partial} = .31$, with an achieved power of 1.00. There was however no significant interaction effect between rhetorical task and L2 proficiency, $F(3, 239) = 0.19$, $p = .90$, $\eta^2_{partial} = .002$, with an achieved power of .09. The results suggest that the lack of relationship between rhetorical task and L2 writing scores for the sample is not dependent on the writers' L2 English proficiency levels.

Table 4.2

Means and Standard Deviations for L2 Writing Scores by Rhetorical Task and L2 Proficiency

	Lower Proficiency			Higher Proficiency		
Group	*n*	*M*	*SD*	*n*	*M*	*SD*
Narrative	26	2.50	0.65	35	3.27	0.66
Expository	32	2.60	0.69	30	3.43	0.63
Expo-Argu	35	2.63	0.64	26	3.46	0.58
Argumentative	34	2.65	0.59	29	3.59	0.56
total	127	2.60	0.64	120	3.43	0.62

Since there were no outliers for the writing scores or the cloze test scores based on *z* scores for each rhetorical task, no separate analyses were conducted with outliers' data removed.

4.1.2 Effects of rhetorical task on CAF of L2 production

A one-way MANOVA test was performed to see the effects of rhetorical task on linguistic accuracy, writing fluency, lexical complexity, and syntactic complexity, with the latter as the dependent variables and rhetorical task as the independent variable. Since the Box's M test showed unequal covariance matrices of the dependent variables, the Pillai's Trace test was used for the MANOVA analysis, instead of the Wilks' Lambda test, as recommended by Meyers, Gamst, and Guarino (2013) and Warner (2013). The Pillai's Trace test result was statistically significant, showing that rhetorical task had a significant effect on the dependent variables at the omnibus level, Pillai's Trace = .57, $F(39, 699) = 4.20$, $p < .01$, $\eta^2_{partial} = .19$, with an achieved power of 1.00.

Table 4.3 displays the descriptive statistics and the univariate analysis results for each dependent variable, with the α level adjusted based on the Holm procedure and the overall α level set at 0.05, since multiple tests were conducted. As can been seen, rhetorical task did not have a significant effect on the linguistic accuracy or fluency of the essays produced by the writers. Rhetorical task also did not have a significant effect on lexical diversity and lexical sophistication of the L2 English essays produced. A significant effect was observed for lexical density; post-hoc pair-wise Tukey tests showed that lexical density of the expository essays was significant higher than that of the essays for all the other three rhetorical tasks, lexical density of the narrative essays was significantly lower than that of the essays for all the other three rhetorical tasks, and lexical density of the expo-argumentative and argumentative essays did not differ.

The descriptive statistics for the different dimensions of syntactic complexity, as displayed in Table 4.3, show an interesting trend: syntactic complexity for the argumentative essays was always the highest, except at the noun-phrase complexity level. The univariate analyses revealed overall significant differences for the following syntactic complexity levels: overall sentence complexity, overall T-unit complexity, overall clause complexity, non-finite subordination, and phrasal coordination. What follows is a summary of the pair-wise Tukey test results for these overall significant differences; only significant comparisons are pointed out, and all the other comparisons were non-significant. Based on pair-wise comparisons, syntactic complexity at the global complexity level–overall sentence complexity as measured by mean length of sentence and overall T-unit complexity as measured by mean length of T-unit--was significantly higher for the argumentative essays, in comparison to syntactic complexity at these global levels found in the essays written for all the other three rhetorical tasks. Overall clause complexity, as measured by mean length of clause, was significantly lower in the narrative essays in comparison to that found in the expository and the argumentative essays. Non-finite subordination, as measured by nonfinite elements per clause, was significantly lower for the expo-argumentative essays in comparison to that for the essays for all the other three rhetorical tasks. Phrasal elaboration through coordinate phrases, as measured by coordinate phrases per verb phrase, was significantly lower for the narrative essays, in comparison to that for the essays for all the other three rhetorical tasks. Note that the significantly lower coordinate phrases per verb phrase for the narrative essays was also the main contributor to the significantly lower mean length of clause for those essays, and this phrasal level of syntactic complexity set the narrative essays apart from all the other essay types. Finally, across the four rhetorical tasks,

Table 4.3

Accuracy, Fluency, Lexical Complexity, and Syntactic Complexity by Rhetorical Task

Construct/ Sub-construct	Measure	Narrative	Expository	Expo-Argu	Argumentative	*F*	*p*	$\eta^2_{partial}$
Accuracy	errors per 100 words	2.78 (1.75)	3.17 (1.97)	3.06 (1.85)	2.48 (1.39)	1.94	0.12	0.02
Fluency	number of words per essay	198.85 (53.70)	204.69 (63.20)	224.21 (57.10)	214.03 (49.98)	2.39	0.07	0.03
Lexical diversity	vocd D	70.51 (17.99)	69.04 (16.57)	73.50 (17.93)	66.64 (17.78)	1.65	0.18	0.02
Lexical sophistication	proportion of sophisticated word types	0.11 (0.04)	0.12 (0.04)	0.12 (0.05)	0.11 (0.05)	1.94	0.12	0.02
Lexical density	lexical words/ all words	0.49 (0.04)	0.54 (0.03)	0.51 (0.03)	0.51 (0.03)	17.26	0.00*	0.18
Overall sentence complexity	mean length of sentence	14.45 (3.04)	15.18 (2.79)	14.93 (3.00)	17.26 (4.11)	8.92	0.00*	0.10
Overall T-unit complexity	mean length of T-unit	13.10 (2.50)	14.06 (2.43)	13.85 (2.66)	15.30 (2.94)	7.42	0.00*	0.08
Clausal coordination	T-units per sentence	1.11 (0.13)	1.09 (0.15)	1.08 (0.12)	1.13 (0.20)	1.42	0.24	0.02
Finite subordination	dependent clauses per T-unit	0.42 (0.21)	0.39 (0.20)	0.42 (0.23)	0.46 (0.24)	1.31	0.27	0.02
Overall clause complexity	mean length of clause	9.30 (1.96)	10.27 (2.07)	9.67 (1.51)	10.43 (1.71)	5.11	0.00*	0.06
Non-finite subordination	nonfinite elements per clause	0.40 (0.19)	0.41 (0.15)	0.29 (0.15)	0.42 (0.18)	7.45	0.00*	0.08
Phrasal coordination	coordinate phrases per verb phrase	0.17 (0.10)	0.30 (0.15)	0.31 (0.14)	0.35 (0.13)	20.90	0.00*	0.21
Noun-phrase complexity	complex NP per verb phrase	0.52 (0.22)	0.59 (0.20)	0.59 (0.23)	0.56 (0.20)	1.33	0.26	0.02

* *p* values are significant with Holm procedure adjustment, with overall α level set at 0.05.

no statistical difference was found in syntactic complexity of their essays at the levels of clausal coordination as measured by T-units per sentence, finite clausal subordination as measured by dependent clauses per T-unit, and phrasal elaboration through complex noun phrases as measured by complex noun phrases per verb phrase.

A one-way MANOVA analysis was also conducted when the outliers based on *z* scores of above 3 or below -3 for each variable were removed, and it revealed the same Pillai's Trace statistical testing result, the overall statistical testing results for each dependent variable were also the same with the ones with the outliers, i.e., the ones reported in Table 4.3, and pair-wise comparisons for the overall significant findings, using the Tukey procedure, yielded exactly the same results with the ones with the outliers, as reported above, except that mean length of T-unit for the argumentative essays was no longer significantly larger than that for the expository essays.

In order to examine whether the writers' general L2 proficiency would make a difference in how rhetorical task affects the CAF features, a two-way MANOVA test was conducted, with task and general L2 proficiency based on the cloze test scores as the independent variables and the CAF features as the dependent variables. Table J1 in Appendix J displays the descriptive statistics for all the CAF features dependent variables, with the data divided by rhetorical task and general L2 proficiency. The two-way MANOVA analysis yielded the following Pillai's Trace test results. Rhetorical task had a significant main effect, Pillai's Trace = .60, $F(39, 687) = 4.37$, $p < .01$, $\eta^2_{partial} = .20$, with an achieved power of 1.00. L2 proficiency level had a significant main effect, Pillai's Trace = .40, $F(13, 227) = 11.79$, $p < .01$, $\eta^2_{partial} = .40$, with an achieved power of 1.00. There was however no interaction between rhetorical task and L2 proficiency level, Pillai's Trace = .18, $F(39, 687) = 1.09$, $p = .33$, $\eta^2_{partial} = .06$, with an achieved power of .96. The findings suggest that the observed effects of rhetorical task on the CAF features were not dependent on the writers' general L2 proficiency. When the two-way MANOVA test was run without the outliers based on 3 *SD* above or below the mean of each variable, the same Pillai's Trace significant testing results were found, showing significant main effects of task and L2 proficiency level but no interaction effect between the two.

4.1.3 Predictive power of CAF on L2 writing scores

4.1.3.1 Correlations between the CAF features and writing scores

First, for each of the rhetorical tasks, bivariate correlations were run, using Pearson product-moment correlations, to see the strengths of the relationships between writing scores and each of the CAF variables and among the CAF variables and to inform predictor selection for multiple regression analyses. Table 4.4 below shows the Pearson *r*s between writing scores and each of the CAF variables for the four rhetorical tasks. Tables K1, K2, K3, and K4 in Appendix K provide the correlation matrixes for the four rhetorical tasks, showing the correlations among the CAF variables as well. Statistical testing results with 2-tailed tests are marked in the tables. The magnitudes of the relationships are interpreted as follows: $r = 0.10$ as being small, $r = 0.30$ as being moderate, and $r = 0.50$ as being large (Huck, 2012). However, due to the multiple tests conducted, any significant results at the $p < .05$ level, not reaching $p < .01$ level, are seen as only marginally significant.

As can be observed in Table 4.4, the strengths of the relationship of most of the CAF variables with writing scores differed across the four rhetorical tasks. The relationships that were statistically significant and relatively consistent across the tasks were associated with measures of writing fluency, lexical sophistication, overall sentence complexity, overall T-unit complexity, and noun-phrase (NP) complexity, and the strengths of these relationships are summarized as follows, with the *r* range shown in the parentheses:

- number of words per essay (.53 - .70)
- proportion of sophisticated word types (.37 - .56)
- mean length of sentence (.25 - .40)
- mean length of T-unit (.29 - .49)
- complex NP per verb phrase (.33 - .46)

Table 4.4

Pearson Correlations for Writing Scores and CAF Features for the Rhetorical Tasks

Construct/ Sub-construct	Measure	Narrative	Expository	Expo-Argu	Argument-ative
Accuracy	errors per 100 words	-0.25	-0.44**	-0.35**	-0.43**
Fluency	number of words per essay	0.60**	0.70**	0.62**	0.53**
Lexical diversity	vocd D	0.36**	0.21	0.38**	0.20
Lexical sophistication	proportion of sophisticated word types	0.37**	0.51**	0.55**	0.56**
Lexical density	lexical words/ all words	0.00	-0.20	-0.09	0.23
Overall sentence complexity	mean length of sentence	0.36**	0.25*	0.40**	0.25*
Overall T-unit complexity	mean length of T-unit	0.45**	0.29*	0.49**	0.41**
Clausal coordination	T-units per sentence	-0.10	-0.03	-0.10	-0.11
Finite subordination	dependent clauses per T-unit	0.20	0.12	0.30*	0.13
Overall clause complexity	mean length of clause	0.21	0.31*	0.42**	0.39**
Non-finite subordination	nonfinite elements per clause	-0.16	0.14	0.27*	0.24
Phrasal coordination	coordinate phrases per verb phrase	0.23	0.17	0.10	0.04
Noun-phrase complexity	complex NP per verb phrase	0.33**	0.36*	0.46**	0.42**

** *p* < 0.01, 2-tailed; * *p* < 0.05, 2-tailed

In addition, the accuracy measure and one additional syntactic complexity measure were both significantly correlated with the writing scores on the expository, expo-argumentative, and argumentative tasks, but not the narrative task:

- errors per 100 words (-.35 - -.44)
- mean length of clause (.31 - .42)

Three more variables were significantly correlated with writing scores on the expo-argumentative task only, and these were the measure of lexical diversity and the two syntactic complexity measures tapping into subordination: dependent clauses per T-unit and nonfinite elements per clause, with moderate correlations ranging from .27 to .38. Further, the lexical diversity measure was also significantly correlated with the writing scores on the narrative task, with a moderate correlation found (.36). There were no other significant relationships between the CAF variables and writing scores on these four tasks.

4.1.3.2 All-possible subsets regression results

To examine the predictive power of the CAF features on L2 writing quality scores when they functioned together, all-possible subsets regression analyses were conducted. Five CAF predictors were selected for the all-possible subsets regression analysis for each task, based on construct representation and the correlation results reported above. The five predictors selected were number of words per essay, grammar and usage errors per 100 words, proportion of sophisticated word types, and mean length of T-unit. These represent all the CAF areas examined–accuracy, fluency, lexical complexity and syntactic complexity. Lexical density was dropped from the analysis since no significant correlations were found between this lexical complexity variable and the writing scores for all the four rhetorical tasks, and it was also shown in the previous literature not to correlate with writing scores (e.g., Engber, 1995; Linnarud, 1986;

Lu, 2012). Overall T-unit complexity was selected as the syntactic complexity variable for the regression analyses because it showed a positive, significant, and relatively consistent relationship with writing scores across the four tasks, its correlation with writing scores was in general the highest if examined across the four tasks, and it represents syntactic complexity at a global level. These predictor variables also did not correlate much with each other; tolerance values for each of the measures were all above .10, showing no problem with multicolinearity, i.e., high inter-correlations among predictor variables. The tolerance value of a given predictor is calculated by subtracting from 1.00 the R^2 values for the given predictor being predicted by all the other predictor variables in a regression analysis, thus showing the inter-correlations or shared variance between this predictor and the other predictors. When the tolerance value for a predictor variable is below .10, the predictor is deemed to have high correlations with the other predictors, and procedures such as combining predictors are recommended.

All-possible subsets regressions were conducted with the above five predictors, using the Excel program–Oshima (2013). Table 4.5 below displays the regression results for the four rhetorical tasks. For each task, the five best regression models are displayed, the first row shows the “best” regression model for that task, and the order of the variables in the first row is based on their importance in predicting the scores for that task, with the most important listed the first. As can be observed, the full model (i.e., the model with all the five predictors) was also among the five best for three of the tasks; for the expository task, it was the sixth best, and it is also listed in the table for comparison purposes. The order of presentation of the five best models for each task is based on the Akaike Information Criterion Corrected (AICC) values, with AICC lower being the better. Mallow’s Cp values are also the smallest for the best regression models for the four tasks, supporting the results for the best models using the AICC values. For each task, all the best models presented in the table have a R^2 that is above the adequate R^2 for that task.

Table 4.5
All-possible-subsets Regression for the Rhetorical Tasks

	Regressors	SSE(k)	R^2	Adj R^2	AICC	Cp	k + 1
Narrative	**F, LD, SC**	17.71	0.48	0.46	-66.74	3.19	4.00
	F, LS, LD, SC	17.47	0.49	0.45	-65.18	4.44	5.00
	F, A, LD, SC	17.52	0.49	0.45	-65.01	4.60	5.00
	F, LD	19.30	0.44	0.42	-63.78	6.24	3.00
	F, A, LS, LD, SC	17.33	0.49	0.45	-63.20	6.00	6.00
Expository	**F, A, LS**	15.52	0.58	0.55	-77.19	3.83	4.00
	F, A, LS, LD	15.20	0.58	0.56	-76.08	4.66	5.00
	F, A, LS, SC	15.24	0.58	0.55	-75.92	4.81	5.00
	F, A	16.45	0.55	0.54	-75.83	5.33	3.00
	F, A, SC	15.89	0.57	0.54	-75.70	5.23	4.00
	F, A, LS, LD, SC	15.03	0.59	0.55	-74.35	6.00	6.00
Expo-Argu	**F, SC, LS, A**	15.62	0.52	0.49	-72.01	4.08	5.00
	F, LS, SC	16.57	0.49	0.47	-70.77	5.44	4.00
	F, A, SC	16.69	0.49	0.46	-70.34	5.86	4.00
	F, A, LS	16.87	0.48	0.46	-69.69	6.49	4.00
	F, A, LS, LD, SC	15.60	0.52	0.48	-69.63	6.00	6.00
Argumenta-tive	**LS, F, A**	17.38	0.49	0.47	-72.45	2.27	4.00
	F, A, LS, SC	17.32	0.49	0.46	-70.28	4.10	5.00
	F, A, LS, LD	17.35	0.49	0.46	-70.17	4.20	5.00
	F, LS	19.30	0.44	0.42	-68.13	6.60	3.00
	F, A, LS, LD, SC	17.30	0.49	0.45	-67.94	6.00	6.00

F = fluency (total # of words); A = accuracy (errors per 100 words); LS = lexical sophistication (proportion of sophisticated word types); LD = lexical diversity (vocd D); SC = syntactic complexity (mean length of T-unit)

As can be observed in the table, for each task, the R^2 values for the best models were the same or very close to each other. Compared across the four tasks, the predictive power of the CAF variables on writing scores was the largest for the expository task (R^2 = 0.59 for the full model, R^2 = 0.58 for the “best” model), either using the full model or the “best” model. The predictive power of the CAF variables on writing scores was mostly the same for the other three rhetorical tasks, with R^2 ranging from .49 to .52 for the full model or the "best" model. These findings show that these CAF variables explained approximately half of the variance in writing scores for the different rhetorical tasks, but their explanatory power was slightly higher for the expository task.

The "best" model for each task revealed through the analyses shows that these CAF variables were not equally important in predicting the writing scores for the different tasks, when they functioned together. For all the four tasks, fewer predictors were able to predict the writing scores well, and the fewer predictors varied across the tasks to some degree. As Table 4.5 shows, the "best" model for each task in general consisted of three or maximally four predictor variables, while their predictive power was exactly the same with or only .01 smaller than the R^2 of the full model, showing that the "best" model could predict as well as the full model while maintaining the fewest predictors possible.

What follows is a detailed comparison of the "best" regression models for the four rhetorical tasks. Table 4.6 below lists the predictors in the "best" model for each task, presented in the order of their importance and with their *b* (regression coefficient) and β (standardized regression coefficient) values. In a multiple regression analysis, *b* shows the amount of increase in the dependent variable with one unit of increase of a given predictor when the other predictors are being held constant, and *b* values are usually not comparable among predictors since their units of calculation are often not the same, but since β is standardized b, β values can be used to compare the importance of different predictors in a multiple regression equation, with a higher β meaning being more important. For example, for the "best" model for the narrative task, the *b* value for the fluency measure was 0.007, and the fluency measure was total number of words;

this means that there is an increase of 0.007 points in the writing score if a writer produces one additional word when there is no change in the values in the other two predictors in the model. The *β* value for the fluency measure, in this case, was .48 and was the highest, showing that the fluency measure was the most important predictor of the writing scores among the three predictors in the "best" model for the narrative task.

Table 4.6
"Best" Regression Models and Regression Coefficients b (β) for the Rhetorical Tasks

Narrative	Expository	Expo-Argu	Argumentative
F .007 (.48)	F .007 (.53)	F .005 (.37)	LS 5.65 (.38)
LD .01 (.25)	A -.10 (-.25)	SC .06 (.22)	F .005 (.33)
SC .07 (.23)	LS 3.22 (.19)	LS 3.37 (.22)	A -.13 (-.25)
		A -.07 (-.18)	

F = fluency (total # of words); A = accuracy (errors per 100 words); LS = lexical sophistication (proportion of sophisticated word types); LD = lexical diversity (vocd D); SC = syntactic complexity (mean length of T-unit)

As Table 4.6 demonstrates, the primary difference among the four tasks is that the "best" model for the narrative task is largely distinct from the ones for the expository, expo-argumentative and argumentative tasks. The "best" model for the narrative task consists of, in the order of importance, the measures of fluency, lexical diversity and overall syntactic complexity, with *b* values of .007, .01, and .07 respectively (*β* = .48, .25, .23). As can also be observed through the other four best models for the narrative task (see Table 4.5), the measures of accuracy and lexical sophistication could not add more predictive power when the other three measures were already in the model. The "best" models for the expository, expo-argumentative and argumentative tasks look much more similar to each other, consisting of the measures of fluency, accuracy and lexical sophistication, although the expo-argumentative task also had the measure of overall syntactic complexity in its “best” model. Similarly, for each of these tasks, adding measures of the other variables not in the “best” model could not increase the predictive power of what was already in the model.

The analysis above shows that the narrative task was distinct from the expository, argumentative types of tasks in the fewest CAF predictors that could well predict its writing scores. As can be further observed in Table 4.6, across all the four rhetorical tasks, the measure of fluency was an important predictor, the most important one based on the β (standardized coefficient) values except for the argumentative task. Interestingly, the importance of the fluency measure in predicting writing scores was noticeably higher for the narrative and expository tasks than that for the expo-argumentative and argumentative tasks, and the importance of the fluency measure dropped linearly along the expository, expo-argumentative and argumentative continuum when the cognitive complexity of the tasks increased. Lexical diversity and overall syntactic complexity were important predictors for the narrative task, but accuracy and lexical sophistication were not. Accuracy and lexical sophistication were however important predictors for the expository, expo-argumentative and argumentative tasks, but lexical diversity was not; overall syntactic complexity was also not an important predictor for the expository and argumentative tasks. Compared across the expository, argumentative types of tasks, lexical sophistication stood out as exceptionally important for predicting writing scores for the argumentative task, cognitively the most complex task. Further, as shown above, the *b* values for each of the CAF variables were not exactly the same across these tasks, calling for slightly different models with different regression coefficients for the expository, expo-argumentative and argumentative tasks.

The all-possible-subsets regression analyses were also conducted after the outliers based on *z* scores and then Standardized DfBeta (DFBETAS) and Cook's D were removed. Only based on *z* scores of 3 *SD* above or below the mean for each variable, there were two outliers for the narrative task, two for the expository task, one for the expo-argumentative task, and three for the argumentative task. There were no outliers identified based on Standardized DfBeta (DFBETAS). With Cook's D, six more outliers were identified for the narrative task, two for the expository task, three for the expo-argumentative task, and four for the argumentative task. There did not seem to be anything special about the outliers identified through Cook's D, except that they were found to greatly affect the regression coefficients. For example, one outlying case might have a very low writing quality rating, but it had a very high accuracy score. Such a case could be identified as an outlier based on Cook's D. Although it may not seem well-justified to have these outliers' data removed, the regression findings without the outliers are summarized below.

In general, the regression patterns delineated from the full data set remained the same for the data set without the outliers. There were however several noticeable changes. First, as expected, the R^2 in the best models all improved for the four tasks; for the "best" of the best models, the R^2 values were .64, .69, .60, and .55 for the narrative, expository, expo-argumentative and argumentative tasks respectively, interestingly with more gains in the predictive power as the cognitive complexity of the tasks decreased. The R^2 for the expository task remained the highest however, now approaching .70, and the predictive power of the CAF variables still dropped with increased cognitive complexity along the exposition, expo-argumentation and argumentation dimension. Second, partly the increased R^2 for the narrative task's "best" model had to do with the addition of another CAF variable in the model; the accuracy measure was now part of the “best” model, together with the measures of fluency, syntactic complexity and lexical diversity. Lexical sophistication was still not found to be an important predictor for the narrative task. Further, for the narrative task’s “best” model, the fluency measure became extremely important for the data without the outliers, and the importance of the other CAF variables also changed to some extent (β = .63, .26, -.15, .14 for the measures of fluency, syntactic complexity, accuracy and lexical diversity respectively). The CAF variables in the “best” models for the expository, expo-argumentative, and argumentative tasks remained the same, although, as expected, the regression coefficients changed to some extent. The order of the importance of the CAF variables in the “best” models all changed slightly for the expository, expo-argumentative and argumentative tasks. Without the outliers’ data, the order of importance for the expository task was measures of fluency, lexical sophistication and accuracy (β = .60, .23, -.22 respectively), for the expo-argumentative task – measures of fluency, accuracy, syntactic complexity and lexical sophistication (β = .47, -.28, .19, .17 respectively), and for the argumentative task – measures of lexical sophistication, accuracy and fluency (β = .42, -.30, .29 respectively). The patterns that the fluency measure was the most important for all the tasks except for the argumentative task while lexical sophistication was the most important for the argumentative task remained.

4.1.4 Summary of main findings

In answering research question 1, the study shows that the college-level EFL learners performed equally well on the rhetorical tasks of narration, exposition, expo-argumentation and argumentation, in terms of the writing quality scores granted; the learners’ general L2 proficiency also did not have an effect on the relationship. In answering research question 2, the study reveals that the college-level EFL learners produced essays with the same level of fluency, linguistic accuracy, lexical diversity and lexical sophistication across the four rhetorical tasks examined, they produced essays with significantly greater lexical density for the expository task and significantly lower lexical density for the narrative task, and they generated essays with significantly greater overall sentence and T-unit syntactic complexity for the argumentative task and significantly lower overall clause complexity for the narrative task; the learners' general L2 proficiency did not have an effect on the observed relationships. In answering research question 3, the study indicates that the CAF variables explained approximately half of the variance in the writing scores on all the four rhetorical tasks; the narrative task was largely distinct from the expository, argumentative types of tasks in having the measures of fluency, lexical diversity and global syntactic complexity at the T-unit level as its predictors in its "best" regression model, while the other three tasks primarily had the measures of fluency, accuracy and lexical sophistication as the predictors in their "best" models, with the expo-argumentative task also including the measure of global syntactic complexity at the T-unit level; the exact predictive power of each of the CAF variables selected in the "best" models and of these variables collectively also differed across the tasks, with patterns that can be associated with the cognitive complexity of the tasks.

4.2 Discussion

This chapter reported on the results on the cognitive complexity dimension of rhetorical task, concerning its effect on L2 writing quality scores, its effects on CAF features in the L2 writing production, and the predictive power of the CAF features on L2 writing scores for the tasks of different cognitive complexity. The rhetorical tasks examined were narrative, expository, expo-argumentative, and argumentative tasks. These tasks form a cline of increasing cognitive demands in terms of the types of thinking required (Hale et al., 1996; Moffett, 1968; Weigle,

2002) and whether reasoning is required and how much reasoning is required at the discourse level (Bain, 1967; Brooks & Warren, 1979; Cairns, 1899; Genung, 1900), with the narrative task the least cognitively complex and the argumentative task the most cognitively demanding.

4.2.1 Discussion of the effect of rhetorical task on L2 writing scores

First, it was found that rhetorical task did not have an effect on the writing scores of the participants. This supports the hypothesis regarding the first research question. It was hypothesized that college-level ESL writers would be cognitively mature enough to handle all these rhetorical tasks, thus being able to perform equally well on not only tasks primarily requiring recalling, but also the ones primarily requiring generalizing and reasoning/inferencing. The lack of performance difference among the tasks for the adult ESL learners observed in this study is aligned with previous findings in such inquiries with populations of similar L2 proficiency levels or cognitive maturity (Greenberg, 1981; Lim, 2009; Spaan, 1993). The finding is in general in contrast to the ones reported for younger L1 learners who have been found to be only able to fare well in the more demanding expository and argumentative types of tasks in later school years (e.g., Calman, 1986; Engelhard et al., 1992; Freedman & Pringle, 1984; Kegley, 1986; Prater & Padia, 1983; Prater, 1985; Sachse, 1984).

4.2.2 Discussion of the effects of rhetorical task on CAF of L2 production

The study then examined the effects of increasing the cognitive complexity of tasks along the rhetorical task dimension on CAF features of the L2 writing produced. It was hypothesized that when the cognitive complexity of these tasks increased, accuracy would decrease, fluency would not change, and linguistic complexity would increase. Some of the hypotheses are supported by the study, and others are not. First, fluency, as predicted, did not vary as a function of rhetorical task. These college-level EFL learners produced essays of statistically the same length across the four tasks. Although L1 younger writers have been found to be only able to produce narrative essays of greater length (Beers & Nagy, 2009; San Jose, 1972), these adult ESL writers in this study were able to produce essays of the same length across these tasks varying in cognitive demands. Indeed, these adult ESL writers produced longer essays for the expo-argumentative and argumentative tasks, in comparison to the ones for the narrative and expository tasks, although the differences were not statistically significant. In Greenberg's (1981) study of L1 English writing by college freshmen, the argumentative essays produced were significantly longer than the expo-argumentative essays; such a statistical difference was not borne out in this study. Perhaps the L2 proficiency level of the writers in this study is not high enough for such a statistical difference to be identified.

Linguistic accuracy of the L2 essays as measured by erater did not decrease as the cognitive complexity of the tasks increased along the rhetorical task dimension, which does not support what was predicted. The adult ESL writers produced essays of statistically the same level of linguistic accuracy for the four rhetorical tasks. In fact, rather, an obvious trend observed in this study was that accuracy increased as the reasoning demand got higher along the exposition, expo-argumentation, and argumentation spectrum, a trend more in support of Robinson's Cognition Hypothesis (Robinson, 2001, 2003, 2005, 2007a, 2010); the differences were closer to being significant when analyzed without the outliers ($p = .08$). The prediction of this study for the effect of rhetorical task on linguistic accuracy was primarily based on the observation that younger L1 writers showed greatly weaker control over written language conventions in argumentative essays than those in narrative essays (Pringle & Freedman, 1979) and on Skehan's Trade-off Hypothesis (Skehan, 1992, 1998; Skehan & Foster, 2001) where the author(s) hypothesize lower accuracy when learners perform on cognitively more complex tasks.

However, it appears that for the adult ESL writers in this study who had studied English for 6-7 years and many of whom still remained at the intermediate level of L2 proficiency, their linguistic accuracy level was perhaps stabilized, with some grammatical errors possibly fossilized. The tasks did not have much an effect on the amount of errors demonstrated. Perhaps errors are more representative of L2 learners' interlanguage level, not easily affected by certain inherent task cognitive features.

It is however interesting to take notice of the obvious trend of increased accuracy in the essays when cognitive complexity increased along the expository, expo-argumentative, and argumentative tasks, which lends some level of support to Robinson's Cognition Hypothesis. In general, the number of errors in the argumentative essays was much smaller than those in the expository and expo-argumentative essays. Robinson's Cognition Hypothesis regarding the effect of increased reasoning demand on accuracy is primarily based on the assumption that learners pay more attention to the accuracy of forms when the task has greater conceptual and functional demands (e.g., Hulstijn, 1989; Tarone, 1985). Perhaps, learners do have the tendency to push for greater linguistic accuracy when conveying meaning on conceptually and functionally more demanding tasks. In this study, the argumentative task had the greatest reasoning demand and asked the learners to take a stand on whether they agreed or disagreed on an issue that was very close to their daily life and to defend their stand with reasons. Then, perhaps when giving their opinions and supporting their positions on the issue of their concern, the learners tried to be as precise and as accurate as possible in their language use when attempting to make their own meaning and arguments across in the L2. In contrast, the expository task simply asked the learners to make generalizations and to tell what they knew and what they had observed or experienced, without making new meaning, thus posing much lower conceptual and functional demands. The learners, in this case, did not seem to have attended to the accuracy of the linguistic forms they produced as much as when they performed on the argumentative task. The different levels of conceptual and functional demands in the expository and argumentative tasks examined in this study also correspond with Bereiter and Scardamalia's (1987) knowledge-telling vs. knowledge-transforming distinctions. It is possible that when learners are engaged in the act of knowledge transformation, their attention to the accuracy of the linguistic forms they use is greatly enhanced, while the act of knowledge-telling does not invite as high a level of attention to precision and accuracy. Although Bereiter and Scardamalia's framework is primarily used to compare composing strategies of novice and experienced writers, the authors also describe tasks as encouraging knowledge-telling or knowledge-transforming.

As for the effect of increased cognitive complexity along the rhetorical task dimension on linguistic complexity, it was hypothesized that the L2 learners would produce language of greater linguistic complexity when cognitive complexity increased. This hypothesis is partially supported by this study, depending on which sub-constructs of lexical complexity or syntactic complexity are examined. Overall, the study shows that the beneficial effect of higher reasoning demand on lexical complexity, as predicted by the researcher and by Robinson's Cognition Hypothesis, is unfounded or marginal, while its beneficial effect on syntactic complexity, as predicted, is largely supported, particularly at the more global levels of syntactic complexity, when the subject matter of the writing is controlled.

For lexical complexity, the study revealed that there was no statistical difference in lexical diversity and lexical sophistication in the essays on the four rhetorical tasks. No previous studies have made similar comparisons. The current study demonstrated that higher reasoning demand in the expo-argumentative and argumentative tasks did not push the learners to produce lexical more diverse or sophisticated language than when they wrote on the narrative and expository tasks that required no or low reasoning demand at the discourse level. The different types of thinking inherent in the different rhetorical tasks–recalling, generalizing and reasoning/inferencing did not have an effect on the lexical diversity and lexical sophistication features of the ESL essays. These levels of lexical complexity do not seem to be able to be manipulated through cognitive complexity along the rhetorical task dimension.

Lexical density, the third lexical complexity feature examined, however, is shown in this study to be significantly affected by rhetorical task, with the narrative essays the least lexically dense and the expository essays the most lexically dense. Ravid (2004) similarly found significantly lower lexical density in the L1 Hebrew narrative essays produced by seventh and eleventh graders than that in the expo-argumentative essays the students produced, equally showing that there is more use of lexical words in expository and argumentative types of essays than in narrative essays. The current study also revealed that lexical density is the highest for expository essays, in comparison to all the other essay types. Overall, the study supports expository, argumentative types of essays as being more informational than narrative essays, by utilizing more lexical words. The study also suggests that expository essays–knowledge-telling types of expository essays use even more lexical words than expo-argumentative and argumentative essays– knowledge-transforming types of task essays, possibly because the primary function of the former is simply to convey information. The higher reasoning demand along the rhetorical task dimension again does not necessarily push the learners to use lexically denser language; rather, knowledge-telling expository tasks perhaps give the learners the greatest opportunities to produce more lexical words. The finding for lexical density only partially supports the prediction of the study and Robinson’s Cognition Hypothesis.

Finally, as far as syntactic complexity (SC) is concerned, the finding of the study provides some good support for the hypothesis that SC increases as reasoning demand increases along the rhetorical task dimension. In general, global syntactic complexity at the sentence level and the T-unit level was found to be significantly higher for the argumentative task, in comparison to all the other tasks, and overall clause complexity was found to be significantly lower for the narrative task, in comparison to the expository task and the argumentative task. However, the differences at these levels of SC were not significant for all the other pair-wise comparisons for the four tasks. For instance, overall sentence complexity and overall T-unit complexity in the expository and the expo-argumentative essays were not found to be greater than those in the narrative essays. In other words, there was no linear increase of global SC with an increase of reasoning demands along the rhetorical task dimension.

The study finding is in congruence with many previous findings that global SC at the T-unit level is significantly greater in argumentative essays than that in narrative essays across grade levels (Beers & Nagy, 2007; Crowhurst & Piche, 1979; Crowhurst, 1980a; Lu, 2011; San Jose, 1972), and that overall clause complexity is significantly greater in argumentative essays than that in narrative essays for older L1 school children or adult L2 writers (Crowhurst & Piche, 1979; Lu, 2011). The finding of the current study is however incongruent with the previous finding that the amount of finite subordination is significantly higher for argumentative essays than that for narrative essays (Beers & Nagy, 2007; Crowhurst & Piche, 1979; Lu, 2011; San Jose, 1972), although Lu (2011), the only study that also examined adult ESL writing, found some nonsignificant differences with some other measures equally indicating the amount of finite subordination. This difference in the findings may be due to the different writer populations in the studies–younger L1 writers vs. adult L2 writers–or the different tasks used, with the narrative task used in the current study being more academically oriented.

Overall though, through this study, it is interesting to observe that SC features at all the local levels except for noun-phrase complexity were all higher for the argumentative essays; it was when all these local-level SC features added up that the global-level SC features in the argumentative essays turned out to be significantly more prominent in comparison to those for all the other tasks. In performing the act of making an argument where a stand is required, writers seem to need to consider more related ideas and concepts and juxtapose these ideas and concepts through all syntactic means of putting meaning into the meaning-bearing units of sentences or T-units. As Crowhurst (1980b) puts it, "When individuals engage in persuasive or argumentative discourse, they are engaging in an activity which, inherently, requires the logical interrelationship of propositions. This results in T-units which are lengthened by the subordination of clauses and less-than-clausal elements.” (p. 229)

The finding for the effect of rhetorical task on syntactic complexity in general supports Robinson’s Cognition Hypothesis (Robinson, 2001, 2003, 2005, 2007a, 2010), which posits that (higher) reasoning demand pushes the learners to produce syntactically more complex language production. Robinson’s prediction of the beneficial effect of reasoning on syntactic complexity is mainly based on the idea of form-function/concept mappings and the notion that “greater structural complexity tends to accompany greater functional complexity in syntax” (Givon, 1985, p. 1021). The prediction is particularly borne out in the most cognitively demanding task of argumentation, which requires substantial amount of reasoning to support and defend one’s stand. In the meanwhile, it should be noted, as pointed out earlier, the significantly greater structural complexity observed in the argumentative essays is primarily at the global SC levels, only marginally at the local SC levels. The finding shows that it is all the specific SC devices utilized together that accompany the articulation of conceptually more complex matters. The finding also demonstrates the importance of using global SC measures in related inquires.

Taken together, the study's findings as to the effects of increased cognitive demand along the rhetorical task dimension on CAF features lend some support to Robinson's Cognition Hypothesis (Robinson, 2001, 2003, 2005, 2007a, 2010), rather than Skehan's Trade-off Hypothesis (Skehan, 1992, 1998; Skehan & Foster, 2001). Most of all, no trade-off effects among CAF were observed, neither for the least cognitively demanding task of narration, nor for the most cognitively demanding task of argumentation. The learners did not produce language of lower accuracy or fluency when they produced language of greater overall syntactic complexity for the argumentative task. They also did not gain fluency or accuracy when they lowered overall syntactic complexity in their language production in the narrative task. Rather, the higher reasoning demand in the argumentative task promoted production of significantly greater overall syntactic complexity and somewhat more accurate language in the essays. The beneficial effects of reasoning demands on syntactic complexity and accuracy support Robinson's Cognition Hypothesis. However, such a beneficial effect was not much found for lexical complexity, and the predicted decrease of fluency for tasks requiring (more) reasoning also could not find its support in this study. The findings of this study thus partially support Robinson's Cognition Hypothesis.

The above findings thus provide some validity evidence for the resource-directing category in Robinson's framework and support the view that increased complexity along the resource-directing dimensions can enhance attention to form-function/concept mappings and provide learners with opportunities to produce syntactically more complex and linguistically more accurate language (Robinson, 2010) and there are multiple attentional resources (Navon, 1989; Wickens, 1989) so that meaning and form may not be competing for scarce attentional resources as Skehan (1992; 1998) grounds his Trade-off Hypothesis. Meaning and form can rather occur in tandem during learners' performance on certain conceptually and functionally demanding tasks. It should also be pointed out that the reasoning demands examined in the current study are in general senses and are more at the discourse-functional level, while the latest framework for the Cognition Hypothesis (Robinson, 2007a) spells out very specific reasoning concepts/functions, i.e., +/− causal reasoning, +/− intentional reasoning, and +/− spatial reasoning. However, the very few studies examining the reasoning in the latest framework for the Cognition Hypothesis have not found the beneficial effects of reasoning on syntactic complexity and accuracy as reported in this study (see Robinson, 2007b for example). Perhaps, the broader sense of reasoning as the cognitive processes involved in making personal interpretation and judgment about something with the employment of reasons and logic at the discourse-functional level can more truly reflect the demand for form-function/concept mappings, a type of thinking in contrast to generalizing and recalling. Further, most of the existing studies on the effects of cognitive complexity on CAF in the task-based teaching literature are studies of spoken language production. It remains to be seen whether the effects of reasoning observed in the current study can also play out in similar L2 speaking tasks.

4.2.3 Discussion of the predictive power of CAF on L2 writing scores

Finally, the study also examined the predictive power of the CAF features in the essays produced on L2 writing quality scores for tasks of different cognitive complexity along the rhetorical task dimension. To the researcher's best knowledge, no previous studies have done similar inquiries that required systematic investigations of the CAF features and informative multiple regression analyses. The current study, through the use of automated tools, was able to study a number of the CAF variables, and through the use of all-possible-subsets regression, was able to compare the best regression models across the different rhetorical tasks.

The study revealed that across the rhetorical tasks, approximately half of the variance in L2 writing quality scores could be explained by CAF features of the essays. This is a rather big portion of the variance explained, given that writing quality is typically assessed through content/idea development, and organization and coherence, along with language production features, as seen in the TOEFL Independent Writing Task rating rubric that was used in the current study. In the meanwhile, it should be noted that language production features and content/idea development often have an inseparable relationship, since for example a good development of ideas in the eyes of the reader may require accurate language use and adequate length to express the ideas. Thereby, when these language production features are examined together with other key criteria for assessing L2 writing, their predictive values may change to some extent.

Perhaps, more interestingly, it was found that these language production features had the greatest predictive power for the expository task and that the predictive power dropped when the cognitive complexity of the tasks increased along the exposition, expo-argumentation and argumentation dimension. This pattern was even more prominent for the writing fluency measure, and writing fluency was generally more important for narration and exposition than expo-argumentation and argumentation. The patterns delineated seem to suggest that for task types that primarily prompt knowledge-telling types of essays, the ability of the learner to be able to write more, tell more knowledge and give more details matters more in the writing quality scores granted, while for task types that encourage knowledge-transforming, although such an ability remains important, other language-related criteria such as how sophisticated the language use is seems to play a greater role in the writing quality as judged by human raters. Notably, lexical sophistication was found to play a more important role than writing fluency for the scores on the argumentative task in this study, while fluency was the most important CAF predictor for all the other rhetorical tasks. All these findings and interpretations are however associated with writing tasks that were completed in a timed-writing setting with only 30 minutes and for L2 writing assessment purposes; the extent to which these task types can promote knowledge-telling or knowledge-transforming might change in other writing situations.

Additionally, it should be pointed out that the above observations about the most important CAF predictors are most likely not separable from the content and function of the rhetorical tasks. Particularly, the expository task asked the writers to tell information that they were all highly familiar with and to make generalizations based on their personal experiences and observations. It was not surprising that many of the writers produced similar content; the researcher, while checking all the essays, found a good number of the writers stating that there were mainly three ways that college students in China used computers and the Internet: studying, communicating with others and entertainment. Then perhaps, for such a rhetorical purpose and the largely similar content in all the essays, the human raters would very likely grant higher scores if the writer is able to tell a bit more information, give more details and provide some specific examples. In contrast, in the act of making an argument on an issue that requires a stand, the writer’s ability to use more sophisticated vocabulary to make their created meaning and arguments clear, precise and convincing matters more in how human raters judge the quality of the essays. In other words, lexical sophistication plays an essential role in the writer’s ability to get his/her own meaning and arguments cross. It is also interesting to note that expo-argumentation, a category created in between exposition and argumentation in this study, does seem to function differently from the other two rhetorical tasks, since it had four predictor variables in its “best” regression model while the other two tasks only had three, and it seems to be a bridging rhetorical task between exposition and argumentation in requiring a bit of everything to predict its scores.

The study also revealed the narrative task to be more distinct from the expository, argumentative types of tasks in terms of the CAF predictors in the best regression models. In particular, lexical diversity was found to be an important predictor of scores for the narrative task, but not for the other tasks, and lexical sophistication was not found to be very important for the scores on the narrative task, but important for the scores on the other tasks. The nature of narrative tasks being typically constrained in a narrow space of time and physical space for a specific event and of expository and argumentative tasks being relatively more open-ended in terms of time and space may explain the different findings associated with lexical diversity. In narration of an experience being constrained in time and space, lexical diversity seems to be more associated with how much detail the writer is able to give while describing the things, people and actions in the event, since more details will entail more thematic content requiring use of different words. While it may be difficult to expand the time and space of a given event, a writer may be able to go deeper into what is in the limited time and space. In contrast, expository and argumentative types of essays are more open-ended in terms of thematic content, not constrained by time and space of specific events, particularly since they would require generalizing across multiple experiences and reasoning/making inferences based on their experiences and knowledge. In such rhetorical situations, lexical diversity seems to be naturally called for by the tasks, and higher or lower lexical diversity in those essays does not seem to affect the quality of the writing as much as when an experience is being blandly or richly recounted. The finding that lexical sophistication is not as important in effective narrative discourse as in effective expository, argumentative types of discourse may be explained by the fact that everyday, common vocabulary may be more expected in narration than in exposition and argumentation where abstraction and thus lower-frequency words are more expected and desired. Thus, not producing sophisticated and lower-frequency words may not be penalized as much for narrative essays as for expository, argumentative types of essays.

In addition to lexical diversity, global syntactic complexity at the T-unit level also did not show up as being an important predictor for writing scores on the expository and the argumentative tasks, but it was important for the expo-argumentative task for a reason explained above. In general, these findings indicate that both lexical diversity and global syntactic complexity do not play much a role in the writing quality of expository and argumentative essays as judged by human raters, particularly when they interact with other language production features of the essays such as essay length and lexical sophistication. Further, the study revealed that the regression coefficients for the different CAF variables in predicting writing scores on the expository, expo-argumentative, and argumentative tasks were slightly different across the tasks, although they all had fluency, accuracy and lexical sophistication in their “best” regression models. This suggests that the regression equations need to slightly different for these different tasks.

4.2.4 Discussion of the relationships among cognitive complexity, L2 writing scores, and CAF of L2 production

A final observation is made here for the relationships ***among*** the cognitive complexity along the rhetorical task dimension, L2 writing quality, and CAF features of L2 writing. Higher cognitive demands along the rhetorical task dimension seem to have higher demands on higher-order language production features–lexical sophistication in particular, since such higher-order language features may be more needed for expressing meaning for higher-order thinking. Lower cognitive demand along the rhetorical task dimension seems to have lower demands on such higher-order language production features, but rather lower-order production features such as essay length would be highly desirable for these tasks. However, these task demands, although found to have played a big role in how the human raters judged the quality of the essays, did not make a difference in the actual essay features produced across the tasks. For example, although the writers were particularly rewarded for writing longer essays for the expository task, they did not produce longer essays for this task in comparison to the other tasks. Although the writers were particularly rewarded for producing more academic and advanced vocabulary for the argumentative task, they did not produce essays with higher lexical sophistication for this task in comparison to the other tasks. In other words, higher task demands on particular CAF features did not make the writers produce essays that could meet the higher demands.

Two reasons might explain such incongruence. First, the L2 writers in the study may not be fully aware of the specific task demands for the different rhetorical tasks, thereby not producing essays to best meet the expectations for higher quality writing on those tasks. In other words, they may not have developed the necessary knowledge, competency and flexibility in writing on all these tasks. Second, although lexical sophistication, as a higher-order CAF feature, is highly desirable for the argumentative task, lexical sophistication is highly related to the L2 proficiency level of the writers; thereby, even when the task has a higher demand, the writers' production of more academic and advanced vocabulary is highly constrained by their L2 proficiency, thus making no difference in the actual lexical features demonstrated. The above observations show that although certain CAF features could potentially be elicited more of through task features, such elicitations can however be constrained due to lack of task knowledge or inadequate L2 proficiency levels. The observations also show the importance of overall L2 proficiency and writing proficiency in predicting writing scores.

CHAPTER 5: RESULTS AND DISCUSSION FOR THE TOPIC FAMILIARITY DIMENSION

This chapter presents the findings about the cognitive complexity dimension of topic familiarity, answering research questions 4, 5, and 6 listed in Chapter 1. Specifically, it presents the results about the effect of topic familiarity on L2 writing scores for college EFL students, the effects of topic familiarity on accuracy, fluency, and linguistic complexity (i.e., lexical complexity and syntactic complexity) of L2 writing production, and the predictive power of the CAF features on L2 writing scores for the different tasks varying in topic familiarity. Topic familiarity is defined in this study as the amount of direct and explicit knowledge writers presumably have developed about a topic through all kinds of experience, such as having direct personal experiences or observations, conversing or thinking about the topic, and obtaining information about the topic from other sources. The construct was operationalized in this study into three levels: higher familiarity (i.e., personal familiar), medium familiarity (i.e., impersonal familiar), and lower familiarity (i.e., impersonal less familiar) topics. The three tasks used in this study for these three levels were all expo-argumentative tasks, were all on the common, everyday subject matter of the use of computers and the Internet, and were realized through asking the writers to discuss the benefits and possible problems of computers and the Internet in relation to themselves, to university students in general (a group they were belonging to and were familiar with), and to people in poor areas of the world (a group they were less familiar with). A discussion of the results follows, in relation to the hypotheses proposed and previous related studies. Implications of the findings for L2 teaching and assessment are discussed in the Conclusion chapter. For a more efficient presentation of the findings, the labels of higher familiarity, medium familiarity, and lower familiarity tasks will be used most of the times rather than personal familiar, impersonal familiar, and impersonal less familiar tasks, while in the discussion section, the latter set of labeling will be used most of the times in order to compare with previous studies.

5.1 Results

5.1.1 Effect of topic familiarity on L2 writing scores

Table 5.1 below shows the descriptive statistics for the L2 writing scores on the three tasks varying in topic familiarity. Through a one-way ANOVA test, it was found that there was no statistical difference among the group means, $F(2, 183) = .78$, $p = .46$, $\eta^2 = 0.008$, with an achieved power of 0.67 for the statistical analysis. The effect size found for this comparison was negligible, and the achieved power was adequate, showing that even with a larger sample size, a difference is not likely to be found. In terms of the writing quality scores given, the participants performed equally well on the three different tasks varying in how familiar they were with the topics.

Table 5.1
Means and Standard Deviations for L2 Writing Scores by Topic Familiarity (the Full Sample)

Group	*n*	*M*	*SD*
Personal-Familiar	63	2.84	0.69
Impersonal-Familiar	61	2.98	0.74
Impersonal-Less familiar	62	2.96	0.70

Since the task fulfillment rating specified in the Method chapter showed that the writers produced essays on the higher and lower familiarity tasks with varying approaches, the analysis was also conducted with only the essays that fulfilled the task demands of being personal-familiar or impersonal-less familiar, so that the findings can more truly reflect the effect of the cognitive demands of the tasks. For the higher familiarity (personal-familiar) task, 21 writers (33%) approached the task personally, while 33 of the 42 remaining writers approached the task impersonally so that they wrote essays as if they were given the impersonal-familiar task. For the medium familiarity (impersonal-familiar) task, 57 writers (93%) were on task, approaching the task impersonally and writing about them or university students in general, and the rest of the four writers all took different approaches with one writer falling into each of the 2, 3, 4, and 5 rating categories for that task (See Appendix I). Regarding the lower familiarity (impersonal-less familiar) task, 35 writers (56%) clearly approached the task by truly considering the less familiar context, as indicated by the adequate and explicit verbal contextualizations they provided in their essays for their discussion about the less familiar context, and the rest of the 27 writers approached the task with general discussions not evidently or only tangibly specific to the less familiar context and produced content more similar to that for the impersonal-familiar task, thus making the less familiar topic more familiar. Table 5.2 below displays the descriptive statistics for the writing scores of those who were on task for the three familiarity tasks. A one-way ANOVA test showed that there was no statistical difference in the means of the writing scores, $F(2, 110) = 1.02$, $p = .36$, $\eta^2 = 0.02$, with an achieved power of 0.68. The analysis further shows that the participants were able to perform equally well on the tasks varying in how familiar they were with the topics. The English proficiency level of the on-task writers, judged by the cloze test scores, was also not higher than that of the off-task writers for both the higher familiarity and the lower familiarity tasks, based on independent *t*-test results, showing that the reason why some of the writers fulfilled these tasks and the others did not was not associated with the writers' L2 proficiency levels.

Table 5.2

Means and Standard Deviations for L2 Writing Scores by Topic Familiarity (the On-task Sample Only)

Group	*n*	*M*	*SD*
Personal-Familiar	21	3.02	0.55
Impersonal-Familiar	57	2.94	0.72
Impersonal-Less familiar	35	3.15	0.71

The study also examined whether the L2 proficiency level of the writers would make a difference in the studied relationships. Table 5.3 below presents the descriptive statistics for the L2 writing scores on the three familiarity tasks (with the full data set), divided based on the two L2 proficiency levels determined by the cloze test. A two-way ANOVA test was conducted, with task and L2 proficiency level as the independent variables and L2 writing scores as the dependent variable. The test showed that topic familiarity did not have a main effect, $F(2, 180) = 1.30$, $p = .27$, $\eta^2_{partial} = .01$, with an achieved power of .28. The writers' L2 proficiency level had a significant main effect, $F(1, 180) = 55.71$, $p < .01$, $\eta^2_{partial} = .24$, with an achieved power of 1.00. There was however no significant interaction effect between topic familiarity and L2 proficiency, $F(2, 180) = .64$, $p = .53$, $\eta^2_{partial} = .007$, with an achieved power of .16. The results mean that the lack of relationship between topic familiarity and L2 writing scores for the participants did not vary as a function of their L2 English proficiency levels. Since there were no outliers for the writing scores or the cloze test scores based on *z* scores of above 3 or below -3 for each of the topic familiarity tasks, no separate analyses were conducted with outliers' data removed.

Table 5.3

Means and Standard Deviations for L2 Writing Scores by Topic Familiarity and L2 Proficiency

	Lower Proficiency			Higher Proficiency		
Group	*n*	*M*	*SD*	*n*	*M*	*SD*
Personal-Familiar	34	2.54	0.53	29	3.18	0.70
Impersonal-Familiar	35	2.63	0.64	26	3.46	0.58
Impersonal-Less familiar	30	2.66	0.65	32	3.25	0.63
total	99	2.61	0.61	87	3.29	0.64

5.1.2 Effects of topic familiarity on CAF of L2 production

A one-way MANOVA test, using the full data set, was first conducted to examine the effects of topic familiarity on complexity, accuracy, and fluency of the L2 writing production, with the CAF features as the dependent variables and topic familiarity as the independent variable. The Box’s M test for the MANOVA analysis showed unequal covariance matrices of the dependent variables; therefore, the Pillai’s Trace test was used, instead of the Wilks' Lambda test, as recommended by Meyers, Gamst, and Guarino (2013) and Warner (2013). The result of the Pillai’s Trace test was statistically significant, showing that topic familiarity had a significant effect on the dependent variables at the omnibus level, Pillai’s Trace = .45, *F*(26, 344) = 4.20, $p < .01$, $\eta^2_{partial} = .23$, with an achieved power of 1.00.

The descriptive statistics and the univariate analysis results for each dependent variable are shown Table 5.4. Since multiple tests were done, the α level was adjusted based on the Holm procedure, with the overall α level set at 0.05, As Table 5.4 indicates, topic familiarity did not have a significant effect on the linguistic accuracy and the fluency of the essays produced by the participants. Topic familiarity however had a significant effect on all the three dimensions of lexical complexity of the L2 English essays produced. Post-hoc pair-wise Tukey tests showed that lexical diversity and lexical sophistication of the essays on the lower familiarity

Table 5.4

Accuracy, Fluency, Lexical Complexity, and Syntactic Complexity by Topic Familiarity

Construct/ Sub-construct	Measure	Personal-Familiar	Impersonal-Familiar	Impersonal-Less familiar	*F*	*p*	$\eta^2_{partial}$
Accuracy	errors per 100 words	2.71 (1.63)	3.06 (1.85)	2.86 (1.72)	0.62	0.54	0.01
Fluency	number of words per essay	218.11 (50.62)	224.21 (57.10)	219.23 (53.79)	0.22	0.80	0.00
Lexical diversity	vocd D	72.37 (17.74)	73.50 (17.93)	61.04 (18.66)	11.71	0.00*	0.11
Lexical sophistication	proportion of sophisticated word types	0.13 (0.04)	0.12 (0.05)	0.09 (0.04)	8.95	0.00*	0.09
Lexical density	lexical words/ all words	0.49 (0.04)	0.51 (0.03)	0.50 (0.04)	6.03	0.00*	0.06
Overall sentence complexity	mean length of sentence	15.22 (3.18)	14.93 (3.00)	16.38 (4.31)	2.90	0.06	0.03
Overall T-unit complexity	mean length of T-unit	13.93 (2.87)	13.85 (2.66)	14.90 (3.01)	2.59	0.08	0.03
Clausal coordination	T-units per sentence	1.10 (0.08)	1.08 (0.12)	1.10 (0.16)	0.30	0.74	0.00
Finite subordination	dependent clauses per T-unit	0.39 (0.20)	0.42 (0.23)	0.46 (0.25)	1.80	0.17	0.02
Overall clause complexity	mean length of clause	9.78 (2.00)	9.67 (1.51)	9.73 (1.72)	0.06	0.94	0.00
Non-finite subordination	nonfinite elements per clause	0.28 (0.17)	0.29 (0.15)	0.21 (0.16)	3.95	0.02	0.04
Phrasal coordination	coordinate phrases per verb phrase	0.30 (0.16)	0.31 (0.14)	0.33 (0.16)	0.89	0.41	0.01
Noun-phrase complexity	complex NP per verb phrase	0.59 (0.20)	0.59 (0.23)	0.73 (0.25)	8.52	0.00*	0.09

* *p* values are significant with Holm procedure adjustment, with overall α level set at 0.05.

task were significantly lower than those of the essays on the other two more familiar tasks. Lexical diversity and lexical sophistication did not differ for the higher and medium familiarity tasks. Lexical density of the essays on the medium familiarity task was significantly higher than that of the essays on higher and lower familiarity tasks. Lexical density did not differ for the higher and lower familiarity tasks.

Many of the syntactic complexity dimensions, as can be observed in Table 5.4, were not affected by topic familiarity, with all the topics controlled at the rhetorical task level of expo-argumentation. The only significant difference, after the Holm procedure adjustment for the multiple tests, lay in noun-phrase complexity: there were significantly more complex noun phrases per verb phrase in the lower familiarity task essays than those in the other two more familiar task essays. The other three syntactic complexity dimensions that had close-to significance results were non-finite subordination, overall sentence complexity and overall T-unit complexity. Non-finite subordination was a lot lower in the lower familiarity task essays than that in the other two more familiar task essays, while overall sentence complexity and overall T-unit complexity were greatly higher for the lower familiarity task essays than those in the other two familiar task essays.

A one-way MANOVA test was also done when the outliers based on 3 *SD* above or below the mean of each variable were removed, and it revealed the same Pillai's Trace statistical test result, the overall statistical testing results for each dependent variable were also the same with the ones with the outliers, i.e., the ones reported in Table 5.4, and pair-wise comparisons for the overall significant findings, using the Tukey procedure, yielded exactly the same results with the ones with the outliers, as reported above.

A separate one-way MANOVA test and subsequent univariate analyses were also conducted for the on-task essays only, with the essays identified based on the task fulfillment rating. Details of these on-task essays are reported in the last section of this chapter. The analyses for the on-task essays only can more truly reflect the effect of cognitive demands due to topic familiarity on CAF features. The one-way MANOVA test for the on-task essays also showed a significant effect of topic familiarity on the CAF features at the overall level, Pillai's Trace = .59, $F(26, 198) = 3.17$, $p < .01$, $\eta^2_{partial} = .29$, power = 1.00. The descriptive statistics and the univariate analyses results are reported in Table L1 in Appendix L. After the α level adjustment using the Holm procedure, as the Table indicates, the only statistical differences were with lexical density and noun-phrase complexity; follow-up post-hoc Tukey tests showed the same pair-wise comparison results for the two variables as the ones reported above for the full sample. The overall trend observed for the lower lexical sophistication and lower lexical diversity in the essays for the lower familiarity task was however still evident in the results for the on-task essays only, with the results approaching significance. Further, the higher overall sentence complexity and overall T-unit complexity observed for the lower familiarity task essays in the full sample was even more evident in the results for the on-task sample only, with even smaller *p* values revealed. These results further show the effects of topic familiarity on some CAF features, particularly lexical and syntactic complexity.

In order to examine whether writers' general L2 proficiency would make a difference in how topic familiarity affects the CAF features, a two-way MANOVA test was conducted, using the full data set, with task and general L2 proficiency based on the cloze test scores as the independent variables and the CAF features as the dependent variables. Table M1 in Appendix M displays the descriptive statistics for all the CAF dependent variables, with the data divided by topic familiarity and general L2 proficiency. The two-way MANOVA analysis yielded the following Pillai's Trace test results. Topic familiarity had a significant main effect, Pillai's Trace = .46, $F(26, 338) = 3.88$, $p < .01$, $\eta^2_{partial} = .23$, with an achieved power of 1.00. L2 proficiency level had a significant main effect, Pillai's Trace = .34, $F(13, 168) = 6.70$, $p < .01$, $\eta^2_{partial} = .34$, with an achieved power of 1.00. There was however no interaction between topic familiarity and L2 proficiency level, Pillai's Trace = .15, $F(26, 338) = 1.08$, $p = .37$, $\eta^2_{partial} = .08$, with an achieved power of .87. The findings suggest that the observed effects of topic familiarity on the CAF features were not dependent on the writers' general L2 proficiency. When the two-way MANOVA test was run without the outliers based on *z* scores of above 3 or below -3 for each variable, the same Pillai's Trace significant testing results were found, indicating significant main effects of task and L2 proficiency level but no interaction effect between the two.

5.1.3 Predictive power of CAF on L2 writing scores

5.1.3.1 Correlations between the CAF features and writing scores

First, Pearson product-moment correlations were run for each of the topic familiarity tasks, to see the strength of relationships between writing scores and each of the CAF variables and among the CAF variables and to inform predictor selection for multiple regression analyses. Table 5.5 below shows the Pearson *r*s between writing scores and each of the CAF variables for the three tasks. The correlation matrixes for each of the three tasks, showing the correlations among the CAF feature variables, can be found in Tables K3, K5, and K6 in Appendix K. Statistical testing results with 2-tailed tests are marked in the tables. The magnitudes of the correlations are interpreted as follows: $r = 0.10$ as being small, $r = 0.30$ as being moderate, and $r = 0.50$ as being large (Huck, 2012). However, since multiple tests were conducted, the significant findings at the *p* < .05 level, not reaching *p* < .01 level, are seen as only marginally significant.

Table 5.5

Pearson Correlations for Writing Scores and CAF Features for the Full Sample for the Topic Familiarity Tasks

Construct/ Sub-construct	Measure	Personal- Familiar	Impersonal- Familiar	Impersonal- Less familiar
Accuracy	errors per 100 words	-0.40**	-0.35**	-0.32*
Fluency	number of words per essay	0.63**	0.62**	0.53**
Lexical diversity	vocd D	0.10	0.38**	-0.09
Lexical sophistication	proportion of sophisticated word types	0.42**	0.55**	0.46**
Lexical density	lexical words/ all words	-0.15	-0.09	0.19
Overall sentence complexity	mean length of sentence	0.18	0.40**	0.05
Overall T-unit complexity	mean length of T-unit	0.23	0.49**	0.21
Clausal coordination	T-units per sentence	-0.13	-0.10	-0.17
Finite subordination	dependent clauses per T-unit	0.06	0.30*	-0.12
Overall clause complexity	mean length of clause	0.28*	0.42**	0.44**
Non-finite subordination	nonfinite elements per clause	0.06	0.27*	0.11
Phrasal coordination	coordinate phrases per verb phrase	0.06	0.10	0.25*
Noun-phrase complexity	complex NP per verb phrase	0.37**	0.46**	0.45**

** *p* < 0.01, 2-tailed; * *p* < 0.05, 2-tailed

As Table 5.5 shows, the strength of relationship of most of the CAF variables with writing scores differed across the three familiarity tasks. The correlations that were statistically significant and relatively consistent across the tasks were associated with measures of accuracy, writing fluency, lexical sophistication, overall clause complexity, and noun-phrase (NP) complexity, and the strengths of these relationships are summarized as follows, with the *r* range shown in the parentheses:

- errors per 100 words (-.32 - -.40)
- number of words per essay (.53 - .63)
- proportion of sophisticated word types (.42 - .55)
- mean length of clause (.28 - .44)
- complex NP per verb phrase (.37 - .46)

In addition, several more variables were significantly correlated with writing scores on the medium familiarity task only, and these were measure of lexical diversity and several measures of syntactic complexity: mean length of sentence, mean length of T-unit, dependent clauses per T-unit, and nonfinite elements per clause, with mostly moderate correlations ranging from .27 to .49. Only one other variable was marginally significantly correlated with writing scores on the lower familiarity task: coordinate phrases per verb phrase (.25). There were no other significant relationships between the CAF variables and scores on the three tasks.

Separate correlation analyses were also conducted for the on-task sample only, to see the strength of relationship between each of the CAF features and writing quality scores of those who fulfilled the tasks as asked. Such analyses can more truly reflect the relationships for tasks of different cognitive complexity along the topic familiarity dimension. However, since the on-task sample was rather small in size for the higher familiarity (*n* = 21) and the lower familiarity (*n* = 35) tasks, the strengths of the relationships found for the two tasks could have been greatly affected. Table 5.6 provides the correlation results for the on-task sample. In general, several significant correlations found for the full sample got a lot stronger with the on-task sample only, going from being moderate to being large. These were associated with the measures of lexical sophistication and overall clause complexity for the higher familiarity task, and the measures of overall clause complexity and noun-phrase complexity for the lower familiarity task. The correlation between the fluency measure and writing scores on the lower familiarity task, however, was weakened for the on-task sample only, dropping from being large for the full sample to being moderate.

Table 5.6

Pearson Correlations for Writing Scores and CAF Features for the On-task Sample for the Topic Familiarity Tasks

Construct/ Sub-Construct	Measure	Personal- Familiar	Impersonal -Familiar	Impersonal -Less familiar
Accuracy	errors per 100 words	-0.36	-0.38**	-0.32
Fluency	number of words per essay	0.69**	0.64**	0.35*
Lexical Diversity	vocd D	0.04	0.38**	-0.11
Lexical Sophistication	proportion of sophisticated word types	0.54*	0.52**	0.46**
Lexical Density	lexical words/ all words	0.13	-0.06	0.35*
Overall Sentence Complexity	mean length of sentence	0.29	0.40**	0.04
Overall T-unit Complexity	mean length of T-unit	0.34	0.49**	0.33
Clausal Coordination	T-units per sentence	-0.14	-0.10	-0.27
Finite Subordination	dependent clauses per T-unit	-0.19	0.26*	-0.27
Overall Clause Complexity	mean length of clause	0.60**	0.45**	0.71**
Non-finite Subordination	nonfinite elements per clause	0.55**	0.30*	0.08
Phrasal Coordination	coordinate phrases per verb phrase	0.21	0.11	0.37*
Noun-Phrase Complexity	complex NP per verb phrase	0.39	0.46**	0.66**

** p < 0.01, 2-tailed; * p < 0.05, 2-tailed

5.1.3.2 All-possible subsets regression results

All-possible subsets regression analyses were conducted to see the predictive power of the CAF features on L2 writing quality scores when they functioned together. Five CAF predictors were selected for the regression analysis for each task, based on construct representation and the correlation results reported above. The five predictors selected were number of words per essay, grammar and usage errors per 100 words, proportion of sophisticated word types, and mean length of clause. These predictors represent all the CAF areas examined–accuracy, fluency, lexical complexity and syntactic complexity. Lexical density was not included in the analysis since the current study did not find significant correlations between this lexical complexity variable and writing scores for all the three topic familiarity tasks, and neither did previous studies (e.g., Engber, 1995; Linnarud, 1986; Lu, 2012). Mean length of clause was selected as the syntactic complexity variable for the regression analyses because it showed a positive, significant, and relatively consistent relationship with writing scores across the three tasks, it represents syntactic complexity at an overall level, and it is also a commonly used syntactic complexity measure in previous writing studies (See Norris & Ortega, 2009). These predictor variables also did not correlate much with each other; tolerance values for each of the measures for each task were all above .10, showing no problem with multicolinearity.

All-possible subsets regressions were conducted with the above five predictors, using the Excel program–Oshima (2013). The regression results for the three topic familiarity tasks are shown in Table 5.7. For each task, the five best regression models are displayed, the first row shows the best regression model for that task, and the order of the variables in the first row is based on their importance in predicting the scores for that task, with the most important listed the first. As can be seen, the full model (i.e., the model with all the five predictors) was also among the five best for the medium and lower familiarity tasks; for the higher familiarity task, it was the 11th best, and it was also listed in the table for comparison purposes. The five best models for each task are ordered and presented based on the Akaike Information Criterion Corrected (AICC) values, with AICC lower being the better. Mallow's Cp values are also the smallest for the best regression models for the three tasks, lending support for the results for the best models using the AICC values. For each task, all the best models presented in the table have a R^2 that is above the adequate R^2 for that task.

Table 5.7

All-possible-subsets Regression for the Topic Familiarity Tasks

	Regressors	SSE(k)	R^2	Adj R^2	AICC	Cp	k + 1
Personal-Familiar	**F, A**	16.23	0.45	0.43	-79.02	2.94	3.00
	F, A, SC	15.67	0.47	0.44	-78.95	2.87	4.00
	F, A, LS	15.80	0.47	0.44	-78.46	3.33	4.00
	F, A, LS, SC	15.46	0.48	0.44	-77.47	4.07	5.00
	F, A, LD	16.22	0.45	0.43	-76.81	4.87	4.00
	F, A, LS, LD, SC	15.44	0.48	0.43	-75.10	6.00	6.00
Impersonal-Familiar	**F, SC, LS, A**	14.88	0.55	0.51	-74.98	4.67	5.00
	F, LS, SC	15.64	0.52	0.50	-74.31	5.52	4.00
	F, A, SC	15.68	0.52	0.50	-74.14	5.68	4.00
	F, A, LS, LD, SC	14.70	0.55	0.51	-73.25	6.00	6.00
	F, A, LD, SC	15.35	0.53	0.50	-73.08	6.43	5.00
Impersonal-Less familiar	**F, SC, LS, A**	15.60	0.48	0.44	-74.47	4.64	5.00
	F, LS, SC	16.40	0.45	0.42	-73.74	5.55	4.00
	F, A, SC	16.47	0.45	0.42	-73.50	5.79	4.00
	F, A, LS, LD, SC	15.42	0.49	0.44	-72.73	6.00	6.00
	F, LS, LD, SC	16.20	0.46	0.42	-72.14	6.82	5.00

F = fluency (total # of words); A = accuracy (errors per 100 words); LS = lexical sophistication (proportion of sophisticated word types); LD = lexical diversity (vocd D); SC = syntactic complexity (mean length of clause)

As Table 5.7 indicates, for each task, the R^2 values for the best models were the same or very close to each other. Compared across the three tasks, the predictive power of the CAF variables on writing scores was slightly higher for the medium familiarity task (R^2 = .55 for both the full model and the "best" model) than that for the higher and lower familiarity tasks (R ranging from .45 to .49 for the full model or the "best" model). These findings show that these CAF variables explained approximately half of the variance in L2 writing scores for the different topic familiarity tasks, but their explanatory power was slightly higher for the medium familiarity task.

An examination of the overall profile of the best regression models for each of the familiarity tasks shows that the best models for the medium and lower familiarity tasks were almost exactly the same, with the exception of a minor difference in the fifth best model for each of the two tasks, while the best models for the higher familiarity task were generally distinct from the ones for the other two tasks. Such differences show the differential importance of some of the CAF variables in predicting writing scores on the higher familiarity task and those on the medium and lower familiarity tasks. Further, the "best" model for the higher familiarity task only consists of two predictors, while the "best" models for the medium and lower familiarity tasks have four predictors; these "best" models were however able to predict as well as the full models.

Below is a detailed comparison of the "best" regression models for the three topic familiarity tasks. Table 5.8 lists the predictors in the "best" model for each task, presented in the order of their importance and with their *b* (regression coefficient) and *β* (standardized regression coefficient) values. The "best" model for the higher familiarity task consists of, in the order of importance, the measures of fluency and accuracy, with *b* values of .008 and -.10 respectively (*β* = .56, -.25). As can also be observed through the other four best models for the higher familiarity task, measures of lexical sophistication, lexical diversity and syntactic complexity basically could not add more predictive power when the fluency and the accuracy measures were already in the model. The "best" models for the medium and lower familiarity tasks were exactly the same, consisting of, in the order of importance, the measures of fluency, syntactic complexity, lexical sophistication, and accuracy. As Table 5.8 displays, the *b* and *β* values for the each of the predictors in the "best" models for these two tasks were also almost the same or were very close across the two tasks, showing their approximately equal importance in predicting scores on the two tasks. The measure of lexical diversity did not turn out to be an important predictor for writing scores across these two tasks. The above analysis shows that the measures of fluency and accuracy were important predictors of scores on the higher familiarity task while all the linguistic complexity measures were not, and that all the CAF predictors in the analysis, except for the measure of lexical diversity, were important predictors of scores on the medium and lower familiarity tasks. With the differences noted, there was however one main similarity in the predictors in the "best" models across the three topic familiarity tasks: the fluency measure was the most important predictor, although its importance for the higher familiarity task was much more pronounced, as can be seen in the larger *b* and *β* values. It can also be observed that the importance of fluency dropped when the cognitive complexity of the tasks increased along the topic familiarity dimension.

Table 5.8

"Best" Regression Models and Regression Coefficients b (β) for the Topic Familiarity Tasks

Personal-Familiar			Impersonal-Familiar			Impersonal- Less familiar		
F	0.008	(0.56)	F	0.006	(0.44)	F	0.005	(0.41)
A	-0.10	(-0.25)	SC	0.13	(0.26)	SC	0.11	(0.26)
			LS	2.96	(0.19)	LS	3.22	(0.19)
			A	-0.06	(-0.16)	A	-0.07	(-0.18)

F = fluency (total # of words); A = accuracy (errors per 100 words); LS = lexical sophistication (proportion of sophisticated word types); LD = lexical diversity (vocd D); SC = syntactic complexity (mean length of clause)

The all-possible-subsets regression analyses were also conducted after the outliers based on *z* scores and then Standardized DfBeta (DFBETAS) and Cook's D were removed. Only based on *z* scores of 3 *SD* above or below the mean for each variable, there were no outliers for the higher and medium familiarity tasks and one outlier for the lower familiarity task. No outliers were identified based on Standardized DfBeta (DFBETAS). Through Cook's D, six outliers were identified for the higher familiarity task, four for the medium familiarity task, and five for the lower familiarity task. The outliers identified through Cook's D did not appear to be unusual cases, except that they greatly affected the regression coefficients. For example, one outlying case might have a very low writing quality rating, but it had a high score on essay length and/or mean length of clause. Such a case could be identified as an outlier based on Cook's D. Although it may not seem well-justified to have these outliers' data removed, the regression findings without the outliers are summarized below.

In general, the regression results for the data set without the outliers were mostly consistent with the results based on the full data set. There were however two main noticeable changes. First, as expected, the R^2 for the best models all improved for the three tasks; for the "best" of the best models, the R^2 values were .57, .60, and .64 for the higher, medium, and lower familiarity tasks respectively. Second, lexical sophistication was no longer part of the "best" regression model for the medium familiarity task, and accuracy was no longer part of "best" model for the lower familiarity task, while the "best" model for the higher familiarity task remained the same, consisting only of the fluency and the accuracy measures. The regression coefficients of the predictors in the "best" models for each task all changed slightly, while the order of importance of the predictors largely remained the same. For the higher familiarity task, the regression coefficients for the fluency and the accuracy measures were .008 (.60) and -.12 (-.31) respectively without the outliers' data. For the medium familiarity task, the regression coefficients for the fluency, syntactic complexity, and accuracy measures were .007 (.56), .14 (.31), and -.09 (-.22) respectively. For the lower familiarity task, the regression coefficients for the fluency, syntactic complexity, and lexical sophistication measures were .007 (.57), .12 (.30) and 2.52 (.17) respectively. The patterns that the fluency measure was the most important predictor for all the tasks and that the lexical diversity measure was not an important predictor of scores across the tasks remained.

Finally, although it is ideal to also conduct the regression analyses with only the on-task sample for each task to study how the cognitive complexity of the tasks might have an impact on the predictive power of the CAF features on L2 writing scores, the small number of on-task cases for the higher and lower familiarity tasks made it impossible to conduct such separate regression analyses. Judging from the correlation results reported in Table 5.6 for the on-task samples only, the regression findings are likely to be somewhat different from the ones found for the full samples.

5.1.4 Summary of main findings

In answering research question 4, the study reveals that the college-level EFL learners performed equally well on the three tasks varying in how familiar they were with the topics – higher familiarity (personal-familiar), medium familiarity (impersonal-familiar), and lower familiarity (impersonal-less familiar) tasks, in terms of the writing quality scores granted; the learners' general L2 proficiency also did not have an effect on the relationship. In answering research question 5, the study demonstrates that the college-level EFL learners produced essays with the same level of fluency, linguistic accuracy, and syntactic complexity (for most of the syntactic complexity dimensions) across the three topic familiarity tasks examined; they produced essays with significantly lower lexical complexity, including lexical diversity, lexical sophistication and lexical density, on the lower familiarity task; they however generated essays with significantly greater noun-phrase complexity and generally much greater overall sentence and T-unit complexity on the lower familiarity task; the learners' general L2 proficiency did not have an effect on the observed relationships. In answering research question 6, the study shows that the CAF variables explained approximately half of the variance in the writing scores on all the three topic familiarity tasks, the fluency measure was the most important CAF predictor of scores across the tasks, the best regression models for the higher familiarity task were more distinct from the ones for the medium and lower familiarity tasks, with the "best" model for the higher familiarity task only composed of the fluency and the accuracy measures while the "best" models for the medium and lower familiarity tasks made up of the fluency, accuracy, lexical sophistication, and overall clause complexity measures but not the lexical diversity measure.

5.2 Discussion

This chapter examined the cognitive complexity dimension of topic familiarity, regarding its effect on L2 writing quality scores, its effects on CAF features in the L2 writing production, and the predictive power of the CAF features on L2 writing scores for tasks of different cognitive complexity. The three levels of such a cognitive complexity dimension examined were higher familiarity (personal-familiar), medium familiarity (impersonal-familiar), and lower familiarity (impersonal-less familiar) tasks, with decreasing amount of direct and explicit knowledge that writers were likely to have already built through experience. Based on the existing writing literature (e.g., Hamp-Lyons & Mathias, 1994) and the task-based language teaching (TBLT) literature (e.g., Skehan, 1998), the three levels of topic familiarity tasks increase in the cognitive demand a writing task imposes on the writer. In addition, according to dual processing theories (Evans, 2010; Evans, 2011; Stanovich, West, & Toplak, 2011), highly compiled knowledge through experience makes cognitive processing autonomous and less effortful, while lack of direct knowledge invites greater cognitive efforts and slower cognitive processing.

5.2.1 Discussion of the effect of topic familiarity on L2 writing scores

First, the current study revealed that topic familiarity did not have an effect on L2 writing quality scores. The writers performed equally well on the tasks that differed in how much direct and explicit knowledge the writers presumably already had on the topics. This finding does not support the study's hypothesis that the writers would obtain higher scores on the medium familiarity (impersonal-familiar) topic. Previous studies however have suggested higher writing scores on impersonal topics than those on personal topics (Hamp-Lyons & Mathias, 1994; Hinkel, 2002; Yu, 2007) and on familiar impersonal topics than those on less familiar impersonal topics (Tedick, 1990). Three reasons may explain the difference of the study finding from previous findings. First, the current study had a tight control over the rhetorical task of the writing tasks and over the subject matter of the writing tasks. All the three topic familiarity writing tasks were expo-argumentative ones, and all of them were on the subject matter of the use of computers and the Internet. It has been suggested in the existing literature that rhetorical task may affect writing quality scores (e.g., Calman, 1986; Prater, 1985; Quellmalz, Capell, & Chou, 1982), so may the subject matter of the writing task (e.g., Calman, 1986; Gabrielson, Gordon, & Engelhard, 1995). The current study had the strength of controlling the potential effects of rhetorical task and subject matter. Second, although the lower familiarity task in the current study is on a less familiar topic, the subject matter of the use of computers and the Internet is itself a familiar one. It is possible that the writers could more easily identify with the subject matter and make inferences on the familiar subject matter about a less familiar context – people in poor areas of the world. Such a process can be difficult if the subject matter is not relatable to the writer's daily life experiences, such as opinions of mercenary soldiers used in Spaan (1993) and Hamp-Lyons & Mathias (1994). Finally, the personal task in the current study can be approached impersonally by taking a "we" stance rather than "I", unlike some personal tasks that probably can only be addressed personally, such as the prompt of "When you go to a party, do you usually talk a lot, or do you prefer to listen? What does this show about your personality?" in Hamp-Lyons & Mathias; this could reduce the comparability of the results of the current study with the previous ones. Further, the analysis conducted with the on-task sample only shows that the mean writing score on the lower familiarity task was actually the highest, although it was not significantly higher than the mean scores on the other two tasks, indicating that high quality essays were produced on the task when the writers really thought of the issues from the perspectives of the less familiar context. One of the raters commented in the post-rating interview that the essays on the lower familiarity task were more interesting to read.

5.2.2 Discussion of the effects of topic familiarity on CAF of L2 production

The study then examined the effects of increased the cognitive complexity of tasks along the topic familiarity dimension on CAF features of the L2 essays produced. It was predicted that when cognitive complexity increases along this dimension, both accuracy and fluency decrease, and the highest linguistic complexity is achieved at the medium familiarity (impersonal-familiar) level, higher than that at the personal-familiar and impersonal-less familiar levels. The findings of the study do not support the hypotheses that accuracy and fluency decrease when college-level

ESL writers write on topics about which they have less direct and explicit knowledge. The writers did not generate essays with significantly greater length or higher accuracy on the personal-familiar topic, and they also did not produce essays with significantly shorter length or lower accuracy on the impersonal-less familiar topic; the writers produced essays with statistically the same length and accuracy for the three topic familiarity topics. Previous studies have found that adult ESL writers produced essays with significantly greater length on more familiar impersonal topics than less familiar impersonal topics (Tedick, 1990) and produced essays with greater length and higher accuracy on personal topics than impersonal topics (Spaan, 1993), suggesting decreased fluency and accuracy with increased task demand along the topic familiarity dimension. The current study however did not show such findings. Again, the current study's tight control over rhetorical task and subject matter, as well as the use of a subject matter close to daily life, may have explained the different findings revealed. Rhetorical task has been found to significantly affect essay length (e.g., Beers & Nagy, 2007; San Jose, 1972; Greenberg, 1981), and subject matter has been found to significantly affect linguistic accuracy (e.g., Clachar, 1999) and essay length (e.g., Yang 2009). Further, the common subject matter of the use of computers and the Internet may have made it easier for the writers to produce essays of adequate length and accuracy even for the impersonal-less familiar topic asking for an examination of the use of computers and the Internet in a less familiar context. The different findings could not be attributed to the varying approaches the writers took for the personal-familiar and impersonal-less familiar tasks, since the on-task samples also did not differ in accuracy and fluency. Overall, the study suggests that adult ESL writers are able to produce L2 essays with the same level of accuracy and fluency when they write on a common, everyday subject matter in relation to themselves, their group, or another group they are less familiar with.

As for the prediction that the highest linguistic complexity is achieved at the impersonal-familiar level, higher than that at the personal-familiar and the impersonal-less familiar levels, the current study finds some support for the hypotheses. In the current study, lexical diversity and lexical sophistication were found to be both significantly higher in the essays on the impersonal-familiar task, than those in the essays on the impersonal-less familiar task, but not higher than those for the personal-familiar task essays; lexical density is significantly higher for the essays on the impersonal-familiar task than that for the essays on the personal-familiar and the impersonal-less familiar tasks. Syntactic complexity, both at the global levels and local levels, was however not higher in the essays on the impersonal-familiar task; rather, noun-phrase complexity was significantly higher for the essays on the lower familiarity, impersonal-less familiar task, and overall sentence complexity and T-unit complexity for those essays were in general much higher, approaching significance.

The finding that lexical diversity is significantly higher for the impersonal-familiar task essays than that for the impersonal-less familiar task essays is congruent with what Yu (2007; 2010) reported for adult ESL writers; no previous studies have examined lexical sophistication of essays on impersonal-familiar tasks and impersonal-less familiar tasks. The study shows that when writers are asked to write about a less familiar context and write on a less familiar topic, the lexical complexity of their essays is greatly weakened. The finding that lexical diversity and lexical sophistication for the impersonal-familiar task essays were not significantly higher than those for the personal-familiar task is different from what Yu (2007; 2010) and Spaan (1993) suggested. This could be due to the fact that many of the writers for the personal-familiar task produced impersonal essays, very much the same with the ones for the impersonal-familiar task; however, since the analysis with the on-task samples only also shows no statistical difference in lexical diversity and lexical sophistication for the two tasks, this interpretation can be ruled out. These different findings yielded in the current study could be attributable to the study design of using the same rhetorical task and the same subject matter across the topic familiarity tasks, as previous studies have reported significant effects of rhetorical task on lexical complexity (e.g., Ravid, 2004; Reid, 1990) and significant effects of subject matter on lexical complexity (e.g., Reynolds, 2002; Yang & Weigle, 2011).

The study's finding that syntactic complexity at the global levels (sentence and T-unit levels) was noticeably a lot higher for the impersonal-less familiar task essays, particularly for the on-task sample only, deserves our attention. The finding is different from Tedick's (1990) observation that overall T-unit complexity is significantly greater for the essays on the more familiar impersonal task used. The current study's use of the same subject matter, a common one, may have made the study finding on syntactic complexity different, as previous studies have found significant effects of subject matter on syntactic complexity (e.g., Crowhurst & Piche, 1979; Yang, Lu, & Weigle, 2012). The much greater overall sentence and T-unit complexity observed for the impersonal-less familiar task essays may be due to the need for the writers to make inferences of the less familiar context based on what they already know and what they have experienced, thus calling for the need to include more propositions and descriptions in the meaning-bearing units of sentences or T-units. The study suggests that when adult L2 writers are writing on a common, everyday subject matter in relation to a less familiar context, the overall syntactic complexity of their writing is greatly enhanced. It was also found through this study that the writers on the impersonal-less familiar task also produced essays with significantly greater noun-phrase complexity. Further studies looking into how this was achieved are needed to explain such a finding.

The findings of the study regarding the effects of increased cognitive complexity along the topic familiarity dimension on CAF features can only provide limited support for Robinson's Cognition Hypothesis (Robinson, 2001; 2003; 2005; 2007a; 2010) and Skehan's Trade-off Hypothesis (Skehan, 1992; 1996; 1998; Skehan & Foster, 2001). Both of these two hypotheses predict lower accuracy, fluency, and complexity when writers perform on tasks where they have lower content knowledge and familiarity, with Skehan predicting some forms of trade-off among CAF. In this study, only lexical complexity – lexical diversity, lexical sophistication, and lexical density decreased when the writers wrote on the lower familiarity topic than when they wrote on the medium familiarity topic. Lexical complexity was however not the highest when the writers wrote on the higher familiarity topic. Thus, there was no linear decrease of lexical complexity when the task complexity increased along the topic familiarity dimension. The study's findings as to the effects of increased cognitive complexity along this dimension on accuracy, fluency, and syntactic complexity do not lend support for Robinson's and Skehan's hypotheses, since both accuracy and fluency did not drop and syntactic complexity (specifically, overall sentence and T-unit complexity and noun-phrase complexity) rather increased when the writers performed on the lower familiarity task. Further, the study did not find any trade-off among CAF for the three topic familiarity tasks, thus providing no support for Skehan's hypothesis about trade-off effects. For instance, lexical complexity was found to be significantly higher for the medium familiarity task, but accuracy or fluency of the essays on this task was not found to be lower. Although the current study only provides very limited support for Robinson's and Skehan's hypotheses on the topic familiarity dimension, it should be pointed out that the current study only examined one type of topic familiarity. As noted earlier, the subject matter used in the current study was very close to the writers' daily life, and the familiarity dimension was realized in this study by requiring the writers to write about a daily-life subject matter in relation to themselves, their group, and another group of people they are less familiar with. The study does investigate one type of topic familiarity; however, its results may not be comparable with the ones based on other types of topic familiarity such as an everyday subject matter vs. a subject matter that is foreign and distant to the writer's daily life.

5.2.3 Discussion of the predictive power of CAF on L2 writing scores

Finally, the study also examined the predictive power of CAF features of the essays produced on L2 writing quality scores for tasks of different cognitive complexity along the topic familiarity dimension. To the researcher's best knowledge, no previous studies have done similar inquiries that required systematic investigations of the CAF features and informative multiple regression analyses. The current study, through the use of automated tools, was able to study a number of the CAF variables, and through the use of all-possible-subsets regression, was able to compare the best regression models across the different tasks.

The study revealed that across the three topic familiarity tasks, CAF features of the essays could explain approximately half of the variance in L2 writing quality scores. This is a rather big portion of the variance explained, given that writing quality is typically assessed through content/idea development, organization, and coherence, along with CAF features such as accuracy and range of vocabulary used, as can be seen in the TOEFL Independent Writing Task rating rubric that was used in the current study. It should be noted however that CAF features and content/idea development often have an inseparable relationship, since for example a good development of ideas in the eyes of the reader may require appropriate and varied vocabulary choices and adequate length to express the ideas. Consequently, when these CAF features are examined together with other key criteria for assessing L2 writing, their predictive values may change to some extent.

It was found that the "best" regression model for the personal-familiar task was more distinct from the ones for the impersonal-familiar and the impersonal-less familiar tasks, with the former only consisting of the fluency and the accuracy measures while the latter composed of the fluency, syntactic complexity, lexical sophistication, and accuracy measures, even though all the tasks were expo-argumentative tasks and were on the same subject matter. The difference observed may be explained by the personal and impersonal distinction; however, what is perhaps more plausible is that there were much greater variations in the responses to the personal-familiar task in terms of the approaches adopted, with almost 2/3 of the essays written as impersonal essays, and those not addressed personally were penalized in terms of the scores given. That is to say, the content and task fulfillment factors may have played a bigger role in predicting the writing quality scores for the personal-familiar task than for the impersonal tasks, making the variables of syntactic complexity and lexical sophistication less important for the scores on the personal task. Nonetheless, how much the writers were able to write in the 30 minutes given, as the fluency measure, remained the most important CAF predictor for scores on all the three topic familiarity tasks. It shows the importance of the sheer amount of writing in writing quality scores, which concurs with Perelman's finding on SAT writing as reported in Winerip (2005).

Further, lexical diversity did not show up as an important predictor of writing scores for any of the three topic familiarity tasks. Although previous studies have reported lexical diversity to be a significant predictor of L2 writing scores (e.g., Yu, 2007; 2010), this study is able to show that when there are other important predictors in the regression analysis, lexical diversity may no longer remain important; further, the correlation analyses in this study show that lexical diversity even did not significantly correlate with writing scores on the personal-familiar and the impersonal-less familiar tasks. These findings cast doubt on the importance of lexical diversity in predicting L2 writing quality scores on expository, argumentative types of tasks.

CHAPTER 6: CONCLUSIONS

In this chapter, a summary of the study and its main findings is first presented. Then implications of the study findings for L2 writing assessment, L2 writing instruction and L2 instruction in general, and theorizing of the cognitive complexity of tasks in the task-based language teaching literature are discussed. The chapter ends with a discussion of the limitations of the study and future study directions.

6.1 Summary of the Study and its Main Findings

The study examined two main cognitive complexity dimensions in L2 writing contexts– rhetorical task and topic familiarity. Drawing on the task-based language teaching literature and the L1 and L2 writing literature, the study examined the effects of these cognitive complexity dimensions on L2 writing quality scores, their effects on linguistic accuracy, writing fluency, and linguistic complexity (including lexical and syntactic complexity) of the L2 production, and the predictive power of these essay features on L2 writing quality scores for tasks of different cognitive complexity. Four levels of rhetorical task were studied: narrative, expository, expo-argumentative, and argumentative tasks, and the topics of these tasks were all familiar to the writers. Three levels of topic familiarity were examined: personal-familiar, impersonal-familiar, and impersonal-less familiar, and all these tasks were controlled at the expo-argumentative rhetorical task level. Six writing prompts were used to study these cognitive complexity dimensions and levels, with one writing prompt shared between these two cognitive complexity dimensions. All the writing prompts were on the subject matter of the use of computers and the Internet, so that potential influences brought by the subject matter were controlled. A total of 375 undergraduate EFL students at a university in Southeast China participated in the study, with each student writing on one of the six tasks and with approximately a total of 60 students writing on each task. The writing task was timed, completed within 30 minutes in each case. The essays were rated by five experienced raters and ESL teachers using the TOEFL iBT Test Independent Writing Rubrics, with half-point ratings added. The essays were also rated on task fulfillment by an experienced ESL teacher and writer and the researcher.

Linguistic accuracy of the essays was assessed through the e-rater engine (version 13.1) of Educational Testing Service, and the measure used was the total number of grammar and usage errors per 100 words. Writing fluency was measured by the total number of words in an essay, and the indices were obtained from the word count function in Microsoft Word. Three sub-constructs of lexical complexity were studies: lexical diversity, lexical sophistication, and lexical density. Lexical diversity was measured by vocd D (Malvern, et al., 2004), and the measure was calculated by the Computerized Language Analysis (CLAN) programs (MacWhinney, 2000). Lexical sophistication was captured through the proportion of sophisticated word types, with sophisticated word types being the ones beyond the first 2,000 most frequent word types determined by the American National Corpus (Reppen, Ide, & Suderman, 2005), and the measure was calculated by the Lexical Complexity Analyzer (Lu, 2012). Lexical density was indicated by the proportion of lexical words, and the measure was calculated by the Lexical Complexity Analyzer (Lu, 2012). Syntactic complexity was measured as a multi-dimensional construct (Norris & Ortega, 2009), with eight interrelated sub-constructs representing syntactic complexity at the global, general levels and the local, specific levels; the measures used for the eight sub-constructs were mean length of sentence, mean length of T-unit, T-units per sentence, mean length of clause, finite dependent clauses per T-unit, non-finite elements per clause, coordinate phrases per verb phrase and complex noun phrases per verb phrase; and these measures were calculated by a computation tool–L2 Syntactic Complexity

Analyzer (Lu, 2010), after some minor adaptations. Analyses of variance were conducted to examine the effects of the cognitive complexity dimensions on L2 writing quality scores and on CAF features. All-possible subsets regression analyses were conducted to investigate the predictive power of the CAF features on L2 writing quality scores for each task, and comparisons of the best regression models were made among the tasks in each of the cognitive complexity dimensions.

The study revealed that cognitive complexity along the rhetorical task dimension did not have an effect on the L2 writing quality scores of the participants'; neither did it have an effect on the linguistic accuracy, writing fluency, lexical diversity, and lexical sophistication of the essays. However, rhetorical task was related to syntactic complexity and lexical density: global syntactic complexity of the argumentative essays was significantly higher than that of the narrative, expository, and expo-argumentative essays, overall clause-level complexity of the narrative essays was significantly lower than that of the essays on the expository and the argumentative tasks, and lexical density was significantly higher for the expository essays and significantly lower for the narrative essays. The regression analyses for the four rhetorical tasks showed that CAF features could explain approximately half of the variance in the scores. Among the CAF features, fluency was the most important predictor of scores for the narrative, expository, and expo-argumentative tasks, but lexical sophistication was the most important in predicting scores for the argumentative task. The "best" regression model for the narrative task was more distinct from the ones for the expository, expo-argumentative, and argumentative tasks, with the former consisting of the fluency, lexical diversity, and global syntactic complexity measures while the latter primarily composed of the fluency, lexical sophistication, and accuracy measures.

Cognitive complexity along the topic familiarity dimension did not have an effect on the L2 writing quality scores of the participants', it also did not have an effect on the linguistic accuracy, writing fluency, and most of the syntactic complexity features of the essays. However, topic familiarity was related to lexical complexity features of the essays: lexical diversity and lexical sophistication were significantly lower in the essays on the lower knowledge, less familiar topic than those in the essays on the two comparatively higher knowledge, more familiar topics, and lexical density of the essays on the less familiar topic was also significantly lower than that in the essays on the impersonal familiar topic. Further, global syntactic complexity of the essays on the less familiar topic was noticeably higher than that of the essays on the two more familiar topics. The regression analyses for the three topic familiarity tasks showed that approximately half of the variance of the scores could be explained by CAF features. Among the CAF features, fluency was the most important CAF predictor of scores across the three tasks. The "best" regression model for the personal-familiar task was distinct from the ones for the impersonal-familiar and the impersonal-less familiar tasks, with the former including the fluency and the accuracy measures whereas the latter having the fluency, accuracy, lexical sophistication, and overall clause-level complexity measures. The results for the multiple-regression analyses for the topic familiarity dimension however need to be interpreted with the understanding that the findings might be different depending on whether the writers fulfilled the tasks as asked, since a number of the writers in this study did not produce personal essays for the personal task and many of them approached the less familiar topic by making it a more familiar one, and the correlation results for the on-task samples only suggest that the predictive power of the CAF features on scores for on-task samples is likely to be different from that for off-task samples.

Through this study, it was revealed that both rhetorical task and topic familiarity, as two important cognitive complexity dimensions, could affect L2 writing performance in some ways. Neither dimension was however found to affect L2 writing scores, although topic familiarity can be said to affect writing scores to some extent since there were a number of writers who did not fulfill the personal and the impersonal less familiar tasks as asked and the writing scores of those writers were negatively affected. Furthermore, neither was found to affect the fluency and accuracy of the writing much, although some beneficial effects on accuracy were found for the argumentative task, the most complex rhetorical task. Both dimensions were however found to affect linguistic complexity–syntactic and lexical complexity. Global syntactic complexity, at the sentence and the T-unit levels, was higher for the essays on the most complex tasks along both of the cognitive complexity dimensions. Rhetorical task was not found to affect lexical diversity or lexical sophistication, but it was found to affect lexical density. Topic familiarity was found to have a great effect on all the lexical complexity features, with significantly lower lexical complexity in the essays on the less familiar topic. The above comparisons show how rhetorical task and topic familiarity uniquely affect L2 writing performance.

The study however was not able to have a fully crossed design, so that potential interaction effects between the two cognitive complexity dimensions were not studied and the results must be understood in view of how each dimension was studied. Specifically, all the rhetorical tasks used in this study were on familiar topics, and rhetorical tasks on less familiar or unfamiliar topics were not examined. There can be writing situations where students are asked to write on any of the rhetorical tasks (e.g., argumentation) on a less familiar topic. In addition, all the topic familiarity tasks used in this study were expo-argumentative tasks, and the other rhetorical tasks along this dimension were not examined. It is likely that students are sometimes asked to write on any of topic familiarity tasks which is of another rhetorical task, e.g., impersonal less familiar narrative tasks. How results pertaining to these other levels and interactions are unknown from the current study and require future investigations.

6.2 Implications

The study has implications for L2 writing assessment, L2 writing instruction, L2 instruction in general, and theorizing of the relationship between the cognitive complexity of tasks and second language performance in CAF areas in the TBLT literature.

6.2.1 Implications for L2 writing assessment

In writing assessment settings, concerns have been primarily placed on the comparability of tasks in terms of writing scores, and there have not been close examinations of textual features such as CAF in considering task comparability. Purpura (2013) however points out the value of studying CAF in language assessment settings, particularly for test validation and formative classroom assessment purposes. The study is able to draw implications for L2 writing assessment, from the vantage points of both writing scores and CAF features.

6.2.1.1 Implications for task selection for L2 writing assessment

The college-level EFL writers performed equally well on all the six tasks varying in the cognitive demands of rhetorical task and topic familiarity, in terms of the writing quality scores granted. This seems to suggest that all these tasks can be used to assess adult ESL learners' L2 English writing, particularly that of the population of the current study–Chinese university students. However, the assessment could probably be only on general L2 writing ability, rather than more specialized L2 writing such as academic writing or business writing which may require specific task types for assessment purposes. For instance, the TOEFL writing tasks are intended to measure the L2 writing proficiency needed to function in college-level academic study in English-speaking countries; a large-scale survey of college-level academic study in the U.S. shows that students in this context primarily only need to write on expository tasks, with some argumentative ones, but not narrative tasks (Hale et al., 1996). In this case, narrative tasks may not be a viable choice for assessing L2 writing, particularly since different language production features may be called for and different assessment criteria may be needed for narrative tasks in contrast to the ones for expository, argumentative types of tasks, as revealed through this study.

Further, task fulfillment ratings revealed that many writers who completed the personal-familiar and the impersonal-less familiar tasks did not complete the tasks as the prompts invited and those writers were penalized to some extent in terms of the scores given; yet almost all the writers for the other four tasks fulfilled the tasks as asked. This invites the question of whether the types of personal-familiar and impersonal-less familiar tasks used in the current study should be used in large-scale standardized assessment settings for a similar population. As it stands, these tasks do not seem to work well for such assessment purposes, since some writers may be penalized for construct-irrelevant factors such as lack of task knowledge and different cultural and cognitive orientations. For the personal task in this study, for instance, the writers might not know that they were expected to address the task personally, not impersonally. The Chinese writers in the study may also tend to use “we” rather than “I” in addressing certain topics in the public discourse, since the Chinese culture is primarily collectivistic (Hofstede, 2001; Oyserman, Coon, & Kemmelmeier, 2002). In contrast, the other tasks seem to be able to do a good job in eliciting responses that more truly represented writers’ L2 writing proficiency, less influenced by construct-irrelevant factors. Such implications are not to be generalized to other types of personal-familiar and impersonal-less familiar tasks though, since for example some types of personal-familiar tasks can only be addressed personally so that the writers cannot easily deviate from task requirements and some types of impersonal-less familiar tasks cannot be approached by making the task a more familiar one if they are on a subject matter that is distant to the writers' everyday life.

Then from the perspectives of task-based language assessment, it is ideal that writers' performance in CAF areas is optimized, so that the writers' interlanguage can be more properly assessed (Skehan, 2001). Equally, from the perspectives of test validity, higher levels of language performance are preferred, since L2 proficiency is one of the main constructs of L2 writing assessments. From such perspectives then, argumentative tasks seem to have a slight advantage in L2 writing assessment, since the task used in the study was found to elicit significantly greater overall syntactic complexity and somewhat more accurate language production. On the other hand, impersonal-less familiar tasks seem to have obvious disadvantages but slight advantages as well, since the essays on the task used in the study were found to be significantly less diverse, less sophisticated, and less dense in lexical choices but greatly more complex in overall syntax at the sentence and the T-unit levels. Given that lexical complexity, particularly lexical sophistication is more important to scores on the impersonal-familiar task in this study than overall sentence- or T-unit complexity is, perhaps impersonal-familiar tasks are less ideal for assessment purposes from this perspective.

6.2.1.2 Implications for rating rubric development, rater training and automated essay scoring

The study used all-possible subsets regression to examine the predictive power of CAF features on L2 writing quality scores for each of the tasks used. The findings from the regression analyses in general point to the need to have more fine-grained rating rubrics for each of the writing task types, to reflect the differential importance of the CAF predictors for each of the tasks. For instance, the study shows that being able to generate as much text and give as much information as possible is highly important for knowledge-telling types of expository tasks and being able to use more advanced or sophisticated vocabulary is exceptionally important for argumentative tasks; then the rating rubrics for those tasks can reflect the importance of such factors by making them clear to the raters. Although the experienced raters and L2 teachers in this study seemed to have applied different criteria for different tasks using the same rubric, it is still potentially important to specify the different criteria in the rubrics. Since without such information, the beginning stage of a rating procedure might see more noise and inconsistencies in the ratings, with raters figuring out the particular features that make essays on a certain task good or poor in quality. Such information is especially important in training of novice raters who may not be adequately aware of or sensitive to task expectations; the specific indications of higher quality writing for a specific task type can be conveyed and discussed with novice raters. Automated Essay Scoring (AES) engines can also benefit from such information to engineer rating that takes into account differential regression coefficients of score predictors for different task types. Having said all these, test designers would also need to decide whether the identified task-specific criteria are important for their test purposes, based on other valid reasons and considerations in their development of rating rubrics. Further, if it is the intention of the test designer to rate all task types using the same criteria, then it would not be reasonable to have different criteria for different task types. In such a case, the implications of the study would point to the importance of rater re-training to eliminate the raters’ tendencies to apply task-specific criteria.

The implications discussed above are however harder to draw for the types of personal-familiar and impersonal-less familiar tasks used in the current study. Although there are clear patterns based on the regression analyses for the full dataset for each of these tasks in this study, the correlation analyses for the on-task samples only suggest that the predictive power of the CAF features and their respective importance may well be different for the on-task samples and the off-task samples for each of these tasks. There were yet many writers who did not complete these tasks as asked. Separate regression analyses for the on-task samples only for the two tasks in this study were not conducted due to the relatively small sample sizes (*n* = 21 and 35 respectively). In general, the correlation findings suggest that the importance of each of the CAF predictors would depend on whether the writers fulfilled these types of tasks as asked. Consequently, such kinds of tasks could pose great challenges to test designers when they develop rating rubrics, since as it appears, different rubrics may be needed for on-task samples and off-task samples. Likewise, novice raters may also have a difficult time distinguishing and applying the different scoring criteria for essays that fulfilled the tasks and those that did not. Automated Essay Scoring (AES) engines will similarly meet challenges in automatically identifying on-task essays and off-task essays for the same writing task and in having different logarithms for them. Perhaps, these findings also suggest that the types of personal-familiar and impersonal-less familiar tasks used in the current study may not be well suited for large-scale standardized L2 writing assessment, due to these complications. This then also points to the importance of piloting test tasks before they are used to identify the ones where students have difficulty in following the task directions and then consider eliminating or revising those tasks.

Another important implication the study has for writing rating rubrics relates to rating criteria connected to lexical complexity and syntactic complexity of essays produced. Influential rating rubrics (e.g., Jacobs, et al., 1981) have the rating criteria of “range of vocabulary” and use of “complex constructions” for lexical and syntactic complexity properties. However, such criteria seem rather ambiguous, particularly in view of the findings of the current study. First, arguably, “range of vocabulary” can have to do with both lexical diversity and lexical sophistication, since with more use of different words comes a wider range of vocabulary and with more use of academic and advanced vocabulary also comes a greater range of vocabulary since use of the most frequent words are often necessitated. The findings of the study however suggest that lexical diversity is not very important for writing quality scores on expository, argumentative types of essays, although it is important for scores on narrative essays, while lexical sophistication is important for scores on expository, argumentative types of essays, but it is not for scores on narrative essays. McNamara, Crossley, and McCarthy (2010) and Yang and Weigle (2011) similarly found lexical sophistication to be a sufficient predictor of writing quality scores on argumentative tasks when studied together with lexical diversity. Based on such observations, the rating criterion of “range of vocabulary” needs to be specified in terms of what it really means. For rubrics for narrative essays, the ability to use more different words can be explicitly specified and emphasized. For rubrics for expository and argumentative essays, the ability to use more academic and advanced vocabulary can be clearly included and stressed. “Range of vocabulary” sounds too ambiguous and does not clearly represent the relationship between lexical complexity and writing quality scores for different task types.

Then, although use of “complex constructions” is evident in influential writing rating rubrics (e.g., Jacobs, et al., 1981), such a criterion for syntactic complexity is equally ambiguous and unclear. The question is what complex constructions are important for writing quality scores, specific ones (e.g., finite subordination) or a constellation of different specific ones? Findings of the current study however suggest that no sub-construct of syntactic complexity, specific ones or overall ones, is likely to be an important predictor of writing quality scores across task types. For example, overall T-unit complexity as measured by mean length of T-unit was not found to be an important predictor of writing scores on the expository and argumentative tasks used in this study, it did not even correlate with writing scores on the personal-familiar and the impersonal-less familiar expo-argumentative tasks, but it was found to be an important predictor of scores on the narrative task and the impersonal-familiar expo-argumentative task. Overall clause complexity as measured by mean length of clause was revealed to be an important predictor of writing scores on the impersonal-familiar and the impersonal-less familiar expo-argumentative tasks, but it was not found to be important in predicting scores on the personal-familiar expo-argumentative task, and the correlation findings also suggest that it is unlikely to be an important predictor for scores on the narrative, expository, and argumentative tasks. Most of the local-level, specific syntactic complexity sub-constructs (e.g., clausal coordination and the amount of nonfinite subordination) were not found to correlate with writing quality scores, and none of them held a significant relationship with writing quality scores across the tasks, with the exception of noun-phrase complexity which requires further investigations.

The above findings about syntactic complexity again suggest that there perhaps need to be specifications of complex constructions that are important for writing quality for each task type. The question then is how to specify such information in rating rubrics. There is no easy answer. Perhaps examining the correlation results between specific, local-level syntactic complexity sub-constructs and writing quality scores for each task type is a good starting point. For instance, for the impersonal-familiar expo-argumentative task, the amount of finite subordination, the amount of non-finite subordination, and noun-phrase complexity were the several specific syntactic complexity sub-constructs that significantly correlated with writing quality scores. Perhaps these complex constructions can be specified as examples in the rating rubrics for this type of writing task. However, this could potentially make the rating task more complex and harder to grasp due to the need to acquire knowledge for the different kinds of complex constructions. Further, complex constructions, when they are used, are to attend certain functions, concepts and meanings. There is potentially an overlap between content and idea development of an essay and the types of complex constructions used. That is to say, specifications of complex constructions important to scores may be redundant if there are clear indications of good content and idea development for a task type. In this way, such a complex task of specifying complex constructions important to scores for each task type and training raters to use such information may not be entirely necessary. It also questions the meaningfulness of the ambiguous and general rating criterion of use of "complex constructions" in rating rubrics. Further investigations are however needed to unpack the relationships between different types of complex constructions and their content/function correlates in relation to writing quality for each task type.

6.2.2 Implications for L2 writing instruction and L2 instruction in general

The findings of the current study have implications for L2 writing instruction and L2 instruction in general. For L2 writing instruction, the study suggests that L2 writing teachers need to help learners gain understandings of task expectations. This includes two aspects. First, L2 writing teachers can help learners develop awareness of criteria that are important for high quality writing for each task type in the eyes of readers or raters. For example, based on this study, how much text a writer is able to generate during timed writing is highly important for writing quality scores across task types; such information could be explicitly conveyed to L2 writers, and perhaps much more importantly, ways that a good length can be produced should be discussed with the learners and be practiced with. A good length does not mean writers should sacrifice clear organization, good content, coherence and so on. Some of the ways to achieve a good essay length can include providing more support and details for main ideas, doing planning before writing, and doing more free-writing on a regular basis to increase general writing fluency. For another example, lexical sophistication, i.e., use of academic and advanced vocabulary, is found to be particularly important for writing quality scores on expository, argumentative types of tasks; such information should also be made clear to L2 writers, and ways to know more academic vocabulary items and to use them in actual writing should be part of a writing course emphasizing these types of writing.

Another aspect that L2 writing teachers can help writers to understand task expectations has to do with fulfillment of task demands. This implication is particularly relevant to the personal-familiar and impersonal-less familiar types of tasks used in the current study. L2 writers shall be aware that if the personal-familiar types of tasks are addressed impersonally and if the impersonal-less familiar types of tasks are treated as familiar tasks and not with sufficient considerations of the less familiar context, their writing quality scores are very likely to be lowered. However, there are cultural issues that might have to be considered and discussed for the personal-familiar types of tasks used in this study. The writers in this study were Chinese university EFL students, while the raters were Americans in the TESOL field in the U.S. In the Chinese culture, addressing certain personal topics impersonally is perhaps a common and socially established practice, since it is a largely collectivistic culture (Hofstede, 2001; Oyserman, Coon, & Kemmelmeier, 2002) where people may be more inclined to speak of “we” rather than “I” in public expo-argumentative discourse; the American culture is however a predominantly individualistic one where individual experiences, perspectives and rights are more emphasized and preferred to be expressed (Hofstede, 2001; Oyserman, Coon, & Kemmelmeier, 2002). Had the personal-familiar essays been rated by Chinese teachers, the results might have been different, which is unknown from the current study. The implication is that such kinds of complications shall be discussed with L2 writers and they may need to adjust their cultural approaches if their writing in high-stake situations is rated by raters from a very different culture, or perhaps the implication is that to accept diverse approaches to the types of personal-familiar tasks by legitimizing different cultural orientations, rating rubrics can specify the acceptability of making certain personal tasks impersonal, or altogether personal tasks should be avoided in high-stakes assessment situations since writers from certain cultures may not feel comfortable writing about themselves in the public discourse.

There are also challenges in providing L2 writing instruction on the impersonal-less familiar types of tasks used in this study. Although there are values in writing on such kinds of tasks, as they call for greater critical thinking skills and the ability to think from others' perspectives and life situations. The challenge for L2 writing teachers then is also to teach such thinking skills, in addition to teaching writing skills and language skills. According to dual processing theories (Evans, 2010; Evans, 2011; Stanovich, West, & Toplak, 2011), it is yet commonplace for individuals to experience cognitive biases through access to and application of short-cut rules and heuristics when confronted with reasoning and rational thinking tasks, showing the difficulty for people to actually do critical thinking and evaluate based on given circumstances. As the findings of the current study demonstrate, a number of college-level EFL students in China perhaps lack the thinking skills to engage with writing tasks that require them to write about less familiar contexts and circumstances. These L2 writers are likely to benefit from instruction on various types of logical thinking, particularly deductive thinking where conclusions are drawn from premises or things they already know to be true. Then these writers would certainly benefit from doing actual writing on different impersonal-less familiar topics, getting feedback from the instructor and their peers, and further learning from revision processes.

The findings of the study also have implications for sequencing of writing tasks in L2 writing instruction and in L2 instruction in general. First, since argumentative tasks are found to have greater demands on linguistic features such as lexical sophistication and overall syntactic complexity than narrative, expository, and expo-argumentative tasks, then for certain thematic content, argumentative tasks can be sequenced after the other task types so that learners' lexical and syntactic repertories can be built up for the linguistically more challenging task type. Similarly, since impersonal-less familiar tasks are found to be more demanding on the linguistic features of lexical complexity and overall syntactic complexity, more familiar tasks on the same thematic content can be sequenced before those tasks so that related lexical items and syntactic structures are activated for the more challenging task type. It is likely that most writing teachers are aware of the importance of sequencing simpler tasks before more complex ones; what this study is able to demonstrate to the teachers is that complex tasks can be linguistically more demanding and thus enrichment of linguistic resources is much needed through performance on simpler tasks on the same thematic content before more complex tasks are used.

Another implication for L2 instruction in general is that argumentative tasks can be particularly helpful for language development, since essays on those tasks are found to have significantly greater overall syntactic complexity and somewhat greater linguistic accuracy and lexical demands are also higher for those tasks. Argumentative tasks should be encouraged to be used in L2 instruction, since they can help learners to further stretch and stabilize their interlanguage, which are part of the goals of task-based language instruction and L2 learning (Skehan, 1992; 1996).

6.2.3 Implications for theorizing of task cognitive complexity in the TBLT literature

6.2.3.1 Implications for linking writing research with TBLT research

The study demonstrated the value of establishing links between different lines of research to inform our understandings of focal questions. In particular, in examinations of the effects of task features on language production features, the task-based language teaching (TBLT) research seems to be able to benefit from L1 and L2 writing research where such questions have also been pursued and CAF features have also been examined. Since few TBLT studies have investigated the effects of task features on CAF of language production in writing tasks and yet the research direction is encouraged (Skehan & Foster, 2001), the large body of existing L1 and L2 writing research can be a fruitful area to draw upon in such inquiries, as the current study demonstrated. In addition to the relatively large number of empirical writing studies that can provide insights and answers to our questions, the writing literature is also rich in terms of writing theories. In particular, theories relating to types of writing such as rhetorical tasks and personal and impersonal tasks examined in this study can be informative. Further, theories relating to writing models (e.g., Hayes, 1996; Grabe & Kaplan, 1996) can be highly relevant as well, since they often provide comprehensive and unique views of factors that can play a role in the writing process and consequently can affect the writing product. These factors are also exactly what the TBLT literature is interested in when studying the effects of task features on CAF. It seems that the theorizing efforts in the TBLT literature along theses lines, particularly in the context of writing tasks, can greatly benefit from the existing L1 and L2 writing research.

6.2.3.2 Implications for theorizing the effects of reasoning and familiarity on CAF in TBLT

One purpose of the current study is to test the competing hypotheses in the TBLT literature regarding the effects of the cognitive complexity of tasks on CAF features. The study does show some support for the beneficial effects of cognitive complexity due to higher reasoning demand in the resource-directing category of Robinson's framework (Robinson, 2001; 2003; 2005; 2007a; 2010), particularly in the area of global syntactic complexity. The main difference between Robinson's hypotheses and those of Skehan's lies in the resource-directing dimension of Robinson's framework and the beneficial effects claimed of increased cognitive complexity on syntactic complexity and linguistic accuracy, as Skehan (1992; 1996; 1998; Skehan & Foster, 2001) predicts lower CAF with all types of cognitive complexity and some sorts of trade-off among CAF. However, what is perhaps very puzzling and revealing through the current study is that the beneficial effect of increased cognitive complexity on global syntactic complexity is also clearly observable for the topic familiarity dimension in this study; Robinson yet groups this cognitive complexity dimension into the resource-dispersing category where he and Skehan have the same hypotheses and thus predict lower syntactic complexity with lower topic familiarity. The question evoked from the findings of the study and thought of by the researcher prior to data collection and data analysis is that when writers are faced with a less familiar topic where their first response might be "I don't know" or "I'm not sure", how the writers can construct responses to address the less familiar topic about which they do not have direct and explicit knowledge or experience. The main option they probably have is to do logical reasoning, particularly deductive reasoning, in which they draw conclusions about what they do not know from premises, i.e., facts or observations that they know to be true. Certainly deductive reasoning and other relevant logical reasoning types are thinking skills and involve greater mental efforts and attention, and if possibly some writers could opt to simply directly apply their own related knowledge and experience to the less familiar topic without doing much of the additional thinking, as some of the writers in the current study may have done. The main argument here is however that when writers are faced with a less familiar topic, what they most likely have to do is to actually engage in a good amount of reasoning. Since reasoning is actually much called for with lower topic familiarity, the argument challenges Robinson's grouping the topic familiarity cognitive complexity dimension into the resource-dispersing category in his framework, not the resource-directing one. Although not much can be said about the other cognitive complexity dimensions in Robinson's framework, the study does indicate that more empirical studies are needed to further validate the resource-directing and resource-dispersing category and their dimensions in Robinson's framework.

The current study however suggests that with higher reasoning demand comes higher global syntactic complexity in language production, as writers find the need to embed more propositions and entities in the meaning-bearing units of T-units or sentences when they put forth their arguments or making inferences about the less known based on what they know. In this way, greater reasoning demand is not only seen in making an argument about what is familiar, but also in making a statement or argument about what is less familiar or unfamiliar. Reasoning in this study means the cognitive processes involved in making a personal interpretation or judgment of something with the employment of reasons and logic. However, the distinction is yet to be made of the effect of greater reasoning demand on lexical complexity, the result of which can depend on whether the writer is reasoning about the familiar or the less familiar. The findings of the current study suggest that when writers are doing reasoning about a familiar topic, lexical complexity of their writing is not affected as compared to when they are only primarily doing generalizing or recalling, but that when writers are doing reasoning about a less familiar topic, lexical complexity of their writing is greatly weakened as compared to when they are writing on a familiar topic. Further, the presence of higher reasoning demands, for familiar or less familiar topics, does not much affect the fluency or the accuracy of L2 writing of college-level writers, as the current study demonstrates.

The above theorizing about the effects of increased reasoning demand on CAF, for both familiar and less familiar topics, however may only be relevant to the writing modality, not the speaking modality, and may only apply to the type of less-familiar topics examined in the current study (i.e., writing about a context writers are less familiar with on a common, everyday subject matter), not other types of less-familiar or unfamiliar topics. In comparison to the writing modality, the speaking modality typically involves more on-time planning and potentially greater cognitive resources due to the pressure to attend to clear and accurate pronunciation of speech, and there are also fewer opportunities for revisions of what is produced in the speaking modality. Together with these basic differences between the writing and the speaking modalities, task conditions can also complicate the relationship between reasoning demand and CAF features. For example, whether L2 writers or speakers are allowed to do planning before their actual language production and whether they are allowed to revise what they produced or to re-perform what they said the first time can all have an effect on the CAF features of their language production (e.g., Bygate, 2001; Ellis & Yuan, 2004; Gilabert, 2005). In this current study, the L2 writers could have their own choice of doing planning before writing and editing after writing in order to simulate the task conditions of a timed writing task in a large-scale standardized assessment; how the findings can be different in other task conditions is unknown.

All these complications due to factors such as the language production modality, the types of less familiar or unfamiliar topics used, and task conditions are what the TBLT literature should heed to in greater details and sophistication in its theorizing about the effects of the cognitive complexity of tasks on CAF features.

6.3 Limitations and future directions

The study has several limitations, and these limitations can also point to future research directions. First, the study participants were Chinese university EFL students. The findings of the study may not be generalizable to other ESL/EFL student populations. It would be interesting to see whether some of the key findings can be borne out in some other student populations. In the meanwhile, given the relatively small number of participants for each task in this study, controlling for the first language of the participants is also a strength of the study, since variations in findings can also be brought by participant background information. Further, with the huge influx of Chinese students in English-speaking countries for undergraduate- and graduate-level studies, the findings of this study can be particularly relevant and useful to those who need to assess the Chinese students’ L2 English writing ability and/or to provide writing instruction to this student population.

Second, although controlling for the subject matter for all the writing tasks is a strength of the study, it can also be seen as a limitation, in that it is unknown whether the findings can be generalizable to other subject matters. It may be possible to replicate the study in the future with another subject matter. Or for future investigations, data from existing large-scale standardized writing assessments can be requested and analyzed by first categorizing the writing tasks into the categories of rhetorical task and degree of topic familiarity and then doing data analyses similar to the ones completed for this study. These future investigations could establish how generalizable the findings are across subject matters.

Thirdly, several of the measures used in the current study are slightly crude and can be further refined in future studies. The accuracy indices–number of grammar and usage errors per 100 words–were obtained from the e-rater engine (version 13.1) of the Educational Testing Service (ETS). Although the engine is believed to do an adequate job, it is probably unable to detect all the grammar and usage errors present in an essay. It however should be able to do a reliable job in identifying the types of errors in all the essays that it is engineered to; thereby it can be said to be doing a consistent job, which human coders may not be able to easily achieve. With these limitations and arguments noted, in the next steps of work, collected essays on selected topics in this study can go through more thorough, manual coding of linguistic errors present, including coding of very specific types of errors. Observations from the current study can be further enlightened.

The study examined syntactic complexity as a multi-dimensional construct, following Norris and Ortega’s (2009) conceptualization and using the existing measures or their adaptations. However, some of the measures used are perhaps still quite crude. For example, the finite subordination and the nonfinite subordination measures can only show overall amount of subordination, finite and nonfinite, and they do not show what grammatical types and functions the subordinate clauses serve. It would be highly valuable to examine these structures more analytically; since different types of subordinate clauses serve different functions, different types of cognitive demands may call for utilization of different types of subordinate clauses, and different subordinate clause types may also have different relationships with writing quality scores. Along this line, work such as Biber, Gray, and Poonpon (2011) and Nippold, Ward-

Lonergan, and Fanning (2005) can be consulted for how to analyze subordinate clauses based on type and function.

Along the line of examining some of the syntactic complexity sub-constructs more closely, in future studies, it will be extremely important and interesting to study how results about some of the syntactic complexity measures are manifested in actual writing, the actual sentences, clauses, nonfinite elements, complex noun phrases and so on produced to see the content correlates of these measures. It would be valuable and interesting to understand from such examinations, for example, what content is associated with significantly lower nonfinite subordination in expo-argumentative essays than that in narrative, expository, and argumentative essays, and what content is related to significantly higher noun-phrase complexity in the lower knowledge, less familiar expo-argumentative essays than that in more familiar expo-argumentative essays and why noun-phrase complexity is very highly correlated with writing quality scores for on-task samples of the less familiar expo-argumentative essays. All of these would require content and function analysis of the complex structures produced in the essays.

Similarly, content and function analysis of lexical complexity features can also be conducted in future studies. For instance, it was hypothesized that the reason why lexical diversity is more important to writing quality scores on narrative tasks than that for expository, argumentative types of tasks is that lexical diversity is associated with giving more details and descriptions about the recounted event for a narrative task. Such an interpretation could be validated by examining the detailedness and vividness of story-telling in the narrative essays collected, in relation to lexical diversity of the essays. Likewise, it would be valuable and interesting to examine the actual, more advanced, sophisticated words used in the argumentative essays collected and to understand from there why they seem to be particularly important to writing scores on argumentative tasks. How are these words potentially related to the meaning-making act in argumentative discourse?

The final limitation of the current study has to do with the writing rating rubric used. The rubric used was the TOEFL iBT Test Independent Writing Rubrics (ETS, 2012), it is a holistic-scoring rubric, and it has its own sets of rating criteria and framework. Had the essays been scored with another rubric, particularly an analytical one, the results could potentially have been different or there could have been opportunities for doing separate analyses for content scores and language scores to unpack the relationships between CAF features and content and between CAF features and language. For future studies, analytical rating rubrics can be considered for examining the relationships explored in the current study.

REFERENCES

American National Corpus (2014). The Open American National Corpus. Retrieved January 20, 2014 from http://www.anc.org

Anderson, L. W., & Krathwohl, D. R. (Eds.). (2001). *A taxonomy for learning, teaching and assessing: A revision of Bloom's Taxonomy of educational objectives: Complete edition*. New York: Longman.

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater_ V.2. *Journal of Technology, Learning, and Assessment, 4*(3). Available from http://www.jtla.org

Bain, A. (1967). *English composition and rhetoric* (2nd ed.). New York: Appleton & company.

Bardovi-Harlig, K. (1992). A second look at T-unit analysis: Reconsidering the sentence. *TESOL Quarterly, 26,* 390–395.

Bardovi-Harlig, K., & Bofman, T. (1989). Attainment of syntactic and morphological accuracy by advanced language learners. *Studies in Second Language Acquisition, 11,* 17–34.

Bartsch, K., & Wellman, H. (1995). *Children talk about the mind.* Oxford: Oxford University Press.

Becker, A., & Carroll, M. (Eds.) (1997). *The acquisition of spatial relations in a second language.* Amsterdam: Benjamins.

Beers, S. F., & Nagy, W. (2009). Syntactic complexity as a predictor of adolescent writing quality: Which measures? Which genre? *Reading and Writing: An Interdisciplinary Journal, 22*, 185-200.

Biber, D. (2006). *University language: A corpus-based study of spoken and written registers.* John Benjamins.

Biber, D., Gray, B., & Poonpon, K. (2011). Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *TESOL Quarterly, 45*, 5-35.

Bloom, B. S. (1956). *Taxonomy of educational objectives: Classifications of educational goal. Handbook I: The cognitive domain.* New York: David McKay Co Inc.

Britton, J., Burgess, T., Martin, N., McLeod, A., & Rosen, H. (1975). *The development of writing abilities (11-18).* London: Macmillan Education.

Brooks, C., & Warren, R. (1979). *Modern rhetoric* (4th ed.). New York: Harcourt.

Brown, J. D. (1980). Relative merits of four methods for scoring cloze tests. *The Modern Language Journal, 64*(3), 311-317.

Brown, J. D. (2002). Do cloze tests work? Or, is it just an illusion? *Second Language Studies, 21*(1), 79-125.

Bygate, M. (2001). Effects of task repetition on the structure and control of oral language. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching pedagogic tasks second language learning and testing* (pp. 23-48). Longman.

Cairns, B. (1899). *Introduction to rhetoric.* Boston: Ginn & company.

Carlisle, R. S. (1989). The writing of Anglo and Hispanic elementary school students in bilingual, submersion, and regular programs. *Studies in Second Language Acquisition, 11,* 257-280.

Carlman, N. (1986). Topic differences on writing tests: How much do they matter? *English Quarterly, 19,* 39-47.

Carlson, S., Bridgeman, B., Camp, R., & Waanders, J. (1985). *The relationship of admission test scores to writing performance of native and nonnative speakers of English* (TOEFL Research Report No. 19). Princeton, NJ: Educational Testing Service.

Clachar, A. (1999). It's not just cognition: The effect of emotion on multiple-level discourse processing in second-language writing. *Language Sciences, 21,* 31–60.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (Rev. ed.). New York: Academic Press.

Coombs, V. (1986). Syntax and communicative strategies in intermediate German composition. *The Modern Language Journal, 70,* 114–124.

Cooper, T. C. (1976). Measuring written syntactic patterns of second language learners of German. *The Journal of Educational Research, 69,* 176–183.

Cooper, T. C. (1981). Sentence combining: An experiment in teaching writing. *The Modern Language Journal, 65,* 158–165.

Corbett, E. (1965). *Classical r*hetoric *for the modern student.* Oxford University Press.

Cristofaro, S. (2003). *Subordination*. Oxford: Oxford University Press.

Cromer, R. (1974). The development of language and cognition: The cognition hypothesis. In B. Foss (Ed.), *New perspectives in child development* (pp. 184-252). Harmondsworth: Penguin.

Crookes, G. (1989). Planning and interlanguage variation. *Studies in Second Language Acquisition, 11,* 367–383.

Crossley, S. A., McNamara, D. S., Weston, J., & McLain Sullivan, S. T. (2011). The development of writing proficiency as a function of grade level: A linguistic analysis. *Written Communication, 28,* 282-311.

Crowhurst, M. (1980a). Syntactic complexity in narration and argument at three grade levels. *Canadian Journal of Education, 5*(1), 6-13.

Crowhurst, M. (1980b). Syntactic complexity and teachers' ratings of narrations and arguments. *Research in the Teaching of English, 13,* 223-231.

Crowhurst, M. (1990). Teaching and learning the writing of persuasive/argumentative discourse. *Canadian Journal of Education, 15*, 348-359.

Crowhurst, M. C., & Piche, G. L. (1979). Audience and mode of discourse effects on syntactic complexity in writing on two grade levels. *Research in the Teaching of English, 13,* 101-109.

Cumming, A. (1989). Writing expertise and language proficiency. *Language Learning, 39,* 81–141.

Daller, H., Van Hout, R., & Treffers-Daller, J. (2003). Lexical richness in the spontaneous speech of bilinguals. *Applied Linguistics, 24*, 197-222.

Devine, J., Railey, K., & Boshoff, P. (1993). The implications of cognitive models in L1 and L2 writing. *Journal of Second Language Writing, 2*, 203-225.

Dewey, J. (1933). *How we think. A restatement of the relation of reflective thinking to the educative process* (Revised ed.). Boston: Heath.

Educational Testing Service (ETS) (2012). TOEFL iBT test independent writing rubrics (scoring standards). Retrieved January 9, 2012, from www.ets.org/Media/Tests/TOEFL/pdf/Writing_Rubrics.pdf

Educational Testing Service (ETS) (2014). About the *e-rater*® scoring engine. Retrieved January 16, 2014, from http://www.ets.org/erater/about

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

Ellis, R. (2009). The differential effects of three types of task planning on the fluency, complexity, and accuracy in L2 oral production. *Applied Linguistics, 30*, 474–509.

Ellis, R., & Yuan, F. (2004). The effects of planning on fluency, complexity and accuracy in second language narrative writing. *Studies in Second Language Acquisition, 26,* 59–84.

Engber, C. (1995). The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second Language Writing, 4*, 139-155.

Engelhard, G., Gordon, B., & Gabrielson, S. (1992). The influences of mode of discourse, experiential demand, and gender on the quality of student writing. *Research in the Teaching of English*, *26*, 315-336.

Evans, J. St. B. T. (2010). *Thinking twice: Two minds in one brain.* Oxford: Oxford University Press.

Evans, J. St. B. T. (2011). Dual-process theories of reasoning: Contemporary issues and developmental applications. *Developmental Review, 31*, 86–102.

Flahive, D., & Snow, B. (1980). Measures of syntactic complexity in evaluating ESL compositions. In J. W. Oller, Jr., & K. Perkins (Eds.), *Research in language testing* (pp. 171–176). Rowley, MA: Newbury House.

Foster, P., & Skehan, P. (1996). The influence of planning and task type on second language performances. *Studies in Second Language Acquisition, 18*, 299–323.

Foster, P., & Skehan, P. (1999). The influence of source of planning and focus of planning on task-based performance. *Language Teaching Research, 3*(3), 215–247.

Foster, P., & Tavakoli, P. (2009). Native speakers and task performance: Comparing effects on complexity, fluency, and lexical diversity. *Language Learning, 59*(4), 866–896.

Frantzen, D. (1995). The effects of grammar supplementation on written accuracy in an intermediate Spanish content course. *The Modern Language Journal, 79*, 329–344.

Freedman, A., & Pringle, I. (1984). Why students can't write arguments. *English in Education, 18,* 73–84.

Gabrielson, S., Gordon, B., & Engelhard, G. (1995). The effects of task choice on the quality of writing obtained in a statewide assessment. *Applied Measurement in Education, 8,* 273–290.

Genung, F. (1900). *The working principles of rhetoric: Examined in their literary relations and illustrated with examples*. Boston: Ginn.

Gilabert, R. (2005). Task complexity and L2 narrative oral production (Unpublished doctoral dissertation), University of Barcelona, Barcelona, Spain.

Gilabert, R. (2007). The simultaneous manipulation of task complexity along planning time and [+/_here-and-now]: Effects on L2 oral production. In M. Garcı´a Mayo (Ed.), *Investigating tasks in formal language learning* (pp. 44–68). Clevedon, UK: Multilingual Matters.

Givon, T. (1985). Function, structure, and language acquisition. In D. I. Slobin (Ed.), *The crosslinguistic study of language acquisition* (Vol. 1, pp.1008-1025). Hillsdale, NJ: Erlbaum.

Givón, T. (1989). *Mind, code and context: Essays in pragmatics*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Givón, T. (2008). *The genesis of syntactic complexity: Diachrony, ontogeny, neuro-cognition, evolution*. Amsterdam: John Benjamins.

Grabe, W., & Kaplan, R. B. (1996). *Theory and practice of writing: An applied linguistic perspective.* London: Longman.

Greenberg, K. L. (1981). *The effects of variations in essay questions on the writing performance of CUNY freshmen.* New York: City University of New York, Instructional Resources Center.

Hale, G., Taylor, C., Bridgeman, B., Carson, J., Kroll, B., & Kantor, R. (1996). *A study of writing tasks assigned in academic degree programs* (TOEFL Research Report No. 54). Princeton, NJ: Educational Testing Service.

Halliday, M. A. K., & Matthiessen, C. (2004). *An introduction to functional grammar* (3rd ed.). London: Arnold.

Hamp-Lyons, L., & Mathias, S. P. (1994). Examining expert judgments of task difficulty on essay tests. *Journal of Second Language Writing, 3*, 85–96.

Hays, J. (1983). *An empirically-derived stage model of the development of analytic writing abilities during the college years: Some illustrative cases* (ERIC Document Reproduction Service No. ED 247 5533).

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

Heatley, A., Nation, I. S. P., & Coxhead, A. (2002). RANGE and FREQUENCY programs. Retrieved January 3, 2013, from http://www.vuw.ac.nz/lals/staff/Paul_Nation

Hedgcock, J., & Lefkowitz, N. (1992). Collaborative oral/aural revision in foreign language writing instruction. *Journal of Second Language Writing, 3,* 255-276.

Henry, K. (1996). Early L2 writing development: A study of autobiographical essays by university-level students of Russian. *The Modern Language Journal, 80,* 309–326.

Hinkel, E. (2002). *Second language writers' text: Linguistic and rhetorical features.* Mahwah, NJ: Erlbaum.

Hinofotis, F. B. (1980). Cloze as an alternative method of ESL placement and proficiency testing. In J. W. Oller, Jr., & K. Perkins (Eds.), *Research in Language Testing* (pp. 121-128). Rowley, MA: Newbury House.

Hoetker, J., & Brossell, G. (1989). The effects of systematic variations in essay topics on the writing performance of college freshman. *College Composition and Communication, 40,* 414-421.

Hofstede, G. (2001). *Culture's consequences: Comparing values, behaviors, institutions, and organizations across nations* (2nd ed.). Thousand Oaks, CA: Sage.

Homburg, T. J. (1984). Holistic evaluation of ESL compositions: Can it be validated objectively? *TESOL Quarterly, 18,* 87–107.

Huberty, C. J. (1989). Problems with stepwise methods-Better alternatives. In B. Thompson (Ed.), *Advances in social science methodology* (Vol. 1, pp. 43-70). Greenwich, CT: JAI.

Huck, S. W. (2012). *Reading statistics and research* (6th ed.). Boston: Pearson Education.

Hulstijn, J. H. (1989). A cognitive view on interlanguage variability. In M. R. Eisenstein (Ed.), *The dynamic interlanguage: Empirical studies in second language variation* (pp. 17–31). New York: Plenum Press.

Hunt, K. W. (1965). *Grammatical structures written at three grade levels*. Champaign, IL: National Council of Teachers of English.

Iwashita, N., McNamara, T., & Elder, C. (2001). Can we predict task difficulty in an oral proficiency test? Exploring the potential of information processing approach to task design. *Language Learning, 51*, 401–436.

Ishikawa, S. (1995). Objective measurement of low-proficiency EFL narrative writing. *Journal of Second Language Writing, 4,* 51–69.

Ishikawa, T. (2007). The effect of manipulating task complexity along the (_Here-and_Now) dimension on L2 written narrative discourse. In C. M. Garcı´a-Mayo (Ed.), *Investigating tasks in formal language learning* (pp. 136–156). Clevedon, UK: Multilingual Matters.

Kameen, P. T. (1979). Syntactic skill and ESL writing quality. In C. Yorio, K. Perkins, & J. Schachter (Eds.). *On TESOL '79: The learner in focus* (pp. 343-364). Washington, D. C.: TESOL.

Kegley, P. H. (1986). The effect of mode of discourse on student writing performance: Implications for policy. *Educational Evaluation and Policy Analysis, 8*(2), 147-154.

Kinneavy, J. L. (1971). *A theory of discourse*. Englewood Cliffs, NJ: Prentice-Hall.

Kohlberg, L. (1983). *The psychology of moral development.* New York: Harper & Row.

Kormos, J. (2011). Task complexity and linguistic and discourse features of narrative writing performance. *Journal of Second Language Writing, 20,* 148-161.

Kuhn, D., & Franklin, S. (2006). The second decade: What develops (and how)? In W. Damon, R. M. Lerner, D. Kuhn, & R. Siegler (Eds.), *Handbook of child psychology (Vol. 2): cognition, perception, and language* (6th ed., pp. 953-993). Hoboken, NJ: Wiley.

Kuiken, F., & Vedder, I. (2007). Task complexity and measures of linguistic performance in L2 writing. *International Review of Applied Linguistics, 45,* 261–284.

Kuiken, F., & Vedder, I. (2008). Cognitive task complexity and written output in Italian and French as a foreign language. *Journal of Second Language Writing, 17,* 48–60.

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied linear statistical models.* (5th ed.). New York: McGraw-Hill.

Langacker, R. W. (2008). *Cognitive grammar: A basic introduction*. Oxford: Oxford University Press.

Larsen-Freeman, D. (1978). An ESL index of development. *TESOL Quarterly, 12,* 439–448.

Laufer, B. (1991). The development of L2 lexis in the expression of the advanced learner. *The Modern Language Journal, 75,* 440–448.

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

Lim, S. G. (2009). *Prompt and rater effects in second language writing performance assessment* (Unpublished Doctoral dissertation). Michigan State University, East Lansing, MI.

Linnarud, M. (1986). *Lexis in composition: A performance analysis of Swedish learners' written English*. Lund: CWK Gleerup.

Long, M. H., & Crookes, G. (1992). Three approaches to task-based syllabus design. *TESOL Quarterly, 26,* 27-56.

Long, M. H., & Crookes, G. (1993). Units of analysis in syllabus design—The case for task. In G. Crookes, & S. M. Gass (Eds.), *Tasks in a pedagogical context: Integrating theory and practice* (pp. 9-54). Clevedon, UK: Multilingual Matters.

Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *TESOL Quarterly, 45*(1), 36-62.

Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal, 96*(2), 190-208.

Lunsford, A. J., Ruszkiewicz, J. J., & Walters, K. (2001). *Everything's an argument, with readings* (2nd ed.). Boston: Bedford/St.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Mallows, C. L. (1973). Some comments on Cp. *Technometrics, 15*, 661-676.

Malvern, D., Richards, B., Chipere, N., & Duran, P. (2004). *Lexical diversity and language development: Quantification and assessment*. Palgrave Macmillan.

McCarthy, P. M., & Jarvis, J. (2007). Vocd: A theoretical and empirical evaluation. *Language Testing, 24*, 459–88.

McNamara, D., Crossley, S., & McCarthy, P. M. (2010). Linguistic features of writing quality. *Written Communication, 27*(1), 57–86.

McNamara, D. S., Louwerse, M. M., & Graesser, A. C. (2002). *Coh-Metrix: Automated cohesion and coherence scores to predict text readability and facilitate comprehension.* Institute for Intelligent Systems, University of Memphis: Memphis, TN.

Mehnert, U. (1998). The effects of different lengths of time for planning on second language performance. *Studies in Second Language Acquisition, 20*, 52–83.

Meyers, L. S., Gamst, G. C. & Guarino, A. J. (2013). *Applied multivariate research: Design and interpretation* (2nd ed.). Newbury Park, CA: Sage.

Moffett, J. (1968). *Teaching the universe of discourse*. Boston, MA: Houghton Mifflin.

Navon, D. (1989). The importance of being visible: On the role of attention in a mind viewed as an anarchic intelligence system. *European Journal of Cognitive Psychology, 1*, 191–238.

Nippold, M. A., Ward-Lonergan, J., & Fanning, J. L. (2005). Persuasive writing in children, adolescents, and adults: A study of syntactic, semantic, and pragmatic development. *Language, Speech, and Hearing Services in Schools, 36*, 125-138.

Nold, E., & Freedman, S. (1977). An analysis of readers' responses to essays. *Research in the Teaching of English, 11*, 164-174.

Norris, J. M., Brown, J.D., Hudson, T., & Yoshioka, J. (1998). *Designing second language performance assessments* (Technical Report No. 18). Hawaii: University of Hawaii Press.

Norris, J. M., & Ortega, L. (2009). Measurement for understanding: An organic approach to investigating complexity, accuracy, and fluency in SLA. *Applied Linguistics, 30,* 555–578.

Ojima, M. (2006). Concept mapping as pre-task planning: A case study of three Japanese ESL writers. *System, 34,* 566–585.

Oller, J. W., & Conrad, C. A. (1971). The cloze technique and ESL proficiency. *Language Learning, 21*(2), 185-195.

Ong, J., & Zhang, L. J. (2010). Effects of task complexity on the fluency and lexical complexity in EFL students' argumentative writing. *Journal of Second Language Writing, 19,* 218-233.

Ortega, L. (1995). *Planning and second language oral performance* (Unpublished MA thesis) University of Hawai'i, Honolulu, HI.

Ortega L. (1999). Planning and focus on form in L2 oral performance. *Studies in Second Language Acquisition, 21,* 109–148.

Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics, 24,* 492-518.

Oshima, C. T. (2013). All possible regressions Excel program. Retrieved January 31, 2014, http://coeweb.gsu.edu/coshima/stat3.htm

Oyserman, D., Coon, H. M., & Kemmelmeier, M. (2002). Rethinking individualism and collectivism: Evaluation of theoretical assumptions and meta-analyses. *Psychological Bulletin, 128*(1), 3–72.

Park, Y. M. (1988). Academic and ethnic background as factors affecting writing performance. In A. C. Purves (Ed.), *Writing across languages and cultures* (pp. 261–272). Beverly Hills, CA: Sage.

Peel, E. A. (1971). *The nature of adolescent judgment*. London: Staples Press.

Perdue, C. (Ed.) (1993). *Adult language acquisition: Crosslinguistic perspectives Vols. 2: The results.* Cambridge: Cambridge University Press.

Piaget, J. (1972). *The psychology of the child*. New York: Basic Books.

Polio, C. (1997). Measures of linguistic accuracy in second language writing research. *Language Learning, 47,* 101-143.

Polio, C. (2001). Research methodology in second language writing research: The case of textbased studies. In T. Silva, & P. K. Matsuda (Eds.), *On second language writing* (pp. 91-115). Lawrence Erlbaum.

Polio, C., Fleck, C., & Leder, N. (1998). "If I only had more time:" ESL learners' changes in linguistic accuracy on essay revisions. *Journal of Second Language Writing, 7,* 43-68.

Prater, D. L. (1985). The effects of modes of discourse, sex of writer, and attitude toward task on writing performance in grade 10. *Educational and Psychological Research, 5,* 241-259.

Prater, D. L. & Padia, W. (1983). Effects of modes of discourse on writing performance on grades four and six. *Research in the Teaching of English, 17,* 127-134.

Pringle, I., & Freedman, A. (1979). *The Carleton writing project, part 1: The writing abilities of a selected sample of grade 7 and 8 students.* Report prepared for the Carleton Board of Education.

Purpura, J. E. (2013). Assessment of grammar. In C. A. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics*. Wiley-Blackwell.

Quellmalz, E., Capell. E, & Chou. C. P. (1982). Effects of discourse and response mode on measurement of writing competence. *Journal of Educational Measurement, 19,* 241-258.

Quinlan, T., Higgins, D., & Wolff, S. (2009). *Evaluating the construct-coverage of the e-rater scoring engine* (ETS Research Report No. RR 09-01). Princeton, NJ: Educational Testing Service.

Raimes, A. (1987). Language proficiency, writing ability, and composing strategies: A study of ESL student writers. *Language Learning, 37,* 439–469.

Ravid, D. (2004). Emergence of linguistic complexity in written expository texts: Evidence from later language acquisition. In D. Ravid & H. Bat-Zeev Shyldkrot (Eds.), *Perspectives on language and language development* (pp. 337–355). Dordrecht: Kluwer.

Ravid, D., & Berman, R. A. (2010). Developing noun phrase complexity at school age: A text-embedded cross-linguistic analysis. *First Language, 30*(1), 3-26.

Read, J. (2000). *Assessing vocabulary*. Oxford, England: Oxford University Press.

Reid, J. (1990). Responding to different topic types: A quantitative analysis from a contrastive rhetoric perspective. In B. Kroll. (Ed.), *Second language writing: Research insights for the classroom* (pp. 191–210). Cambridge, UK: Cambridge University Press.

Reppen, R., Ide, N., & Suderman, K. (2005). *American National Corpus (ANC) Second Release.* Linguistic Data Consortium, Philadelphia.

Reynolds, D. W. (2002, December). Linguistic and cognitive development in the writing of middle-grade English language learners. *Southwest Journal of Linguistics*. Retrieved November 26, 2009, http://findarticles.com/p/articles/mi_hb1440/is_2_21/ai_n28971691/

Robb, T., Ross, S., & Shortreed, I. (1986). Salience of feedback on error and its effect on EFL writing quality. *TESOL Quarterly, 20*, 83-95.

Robinson, P. (1995). Task complexity and second language narrative discourse. *Language Learning*, 45, 99-140.

Robinson P. (2001). Task complexity, cognitive resources and syllabus design: A triadic framework for examining task influences on SLA. In P. Robinson (Ed.), *Cognition and second Language instruction* (pp. 285-316). Cambridge: Cambridge University Press.

Robinson, P. (2003). The Cognition Hypothesis, task design and adult task-based language learning. *Second Language Studies, 21*(2), 45–107.

Robinson, P. (2005) Cognitive complexity and task sequencing: A review of studies in a Componential Framework for second language task design. *International Review of Applied Linguistics in Language Teaching, 43*(1), 1–33.

Robinson, P. (2007a). Criteria for classifying and sequencing pedagogic tasks. In M. Garcı´a Mayo (Ed.), *Investigating tasks in formal language learning* (pp. 7-27). Clevedon, UK: Multilingual Matters.

Robinson, P. (2007b). Task complexity, theory of mind, and intentional reasoning: Effects on speech production, interaction, uptake and perceptions of task difficulty. *International Review of Applied Linguistics, 45*, 193–214.

Robinson, P. (2010). Situating and distributing cognition across task demands: The SSARC model of pedagogic task sequencing. In M. Putz, & L. Sicola (Eds.), *Cognitive processing in second language acquisition: Inside the learner's mind* (pp. 239-264). Amsterdam /Philadelphia PA: John Benjamins.

Robinson, P., Ting, S., & Urwin, J. (1995). Investigating second language task complexity. *RELC Journal, 25,* 62-79.

Robinson, P., & Gilabert, R. (2007). Task complexity, the Cognition Hypothesis and language learning and performance. *International Review of Applied Linguistics, 43,* 161–176.

Sachse, P. P. (1984). Writing assessment in Texas: Practices and problems. *Educational measurement: Issues and practice, spring*, 21-23.

Sangarun, J. (2005). The effects of focusing on meaning and form in strategic planning. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 111-141). Amsterdam: John Benjamins.

San Jose, C. P. M. (1972). *Grammatical structures in four modes of writing at four grade levels* (Unpublished dissertation). Syracuse University, NY.

Sasaki, M., & Hirose, K. (1996). Explanatory variables for EFL students' expository writing. *Language Learning, 46*, 137–168.

Sato, C. (1988). Origins of complex syntax in interlanguage development. *Studies in Second Language Acquisition, 10,* 371-395.

Sato, C. (1990). *The syntax of conversation in interlanguage development.* Tubingen: Gunter Narr.

Shaw, P., & Liu, E. (1998). What develops in the development of second-language writing? *Applied Linguistics, 19,* 225-254.

Skehan, P. (1992). Strategies in second language acquisition. *Thames Valley University Working Papers in English Language Teaching,* No. 1.

Skehan, P. (1996). A framework for the implementation of task based instruction. *Applied Linguistics, 17*(1), 38–62.

Skehan, P. (1998). *A cognitive approach to language learning.* Oxford University Press.

Skehan, P. (2001). Tasks and language performance. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching pedagogic tasks, second language learning, teaching, and testing* (pp. 167-185). Longman.

Skehan, P. and Foster, P. (1997). The influence of planning and post-task activities on accuracy and complexity in task based learning, *Language Teaching Research 1*(3), 185–211.

Skehan, P., & Foster, P. (2001). Cognition and tasks. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 181-203). Cambridge: Cambridge University Press.

Skehan, P. & Foster, F. (2005). Strategic and on-line planning: The influence of surprise information and task time on second language performance. In R. Ellis (ed.), *Planning and task performance in a second language* (pp. 193-216). Amsterdam: John Benjamins.

Smith, C. S. (2003). *Modes of discourse: The local structure of texts*. Cambridge. University Press.

Spaan, M. (1993). The effect of prompt on essay examinations. In D. Douglas, & C. Chapelle (Eds.), *A new decade of language testing research* (pp. 98–122). Alexandria, VA: TESOL.

Stanovich, K. E., West, R. F., & Toplak, M. E., (2011). The complexity of developmental predictions from dual process models. *Developmental Review, 31,* 103–118.

Stevens, J. (2009). *Applied multivariate statistics for the social sciences* (5th ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Talmy, L. (2000*). Toward a cognitive semantics, Vol. 1: Concept structuring systems*. Cambridge, MA: MIT Press.

Tarone, E. (1985). Variability in interlanguage use: A study of style-shifting in morphology and syntax. *Language Learning, 35,* 373-403.

Tarone, E., Downing, B., Cohen, A., Gillette, S., Murie, R., & Dailey, B. (1993). The writing of Southeast AsianAmerican students in secondary school and university. *Journal of Second Language Writing, 2,* 149-172.

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

Van Hout, R., & Vermeer, A. (2007). Comparing measures of lexical richness. In H. Daller, J. Milton, & J. Treffers-Daller (Eds.), *Modelling and assessing vocabulary knowledge* (pp. 93-115). New York: Cambridge University Press.

VanPatten, B. (1990). Attending to content and form in the input: An experiment in consciousness. *Studies in Second Language Acquisition, 12,* 287–301.

VanPatten, B. (1994). Evaluating the role of consciousness in second language acquisition: Terms, linguistic features and research methodology. *AILA Review, 11,* 27-36.

von Stutterheim, C. (1991). Narrative and description; Temporal reference in second language acquisition. In T. Huebner, & C. Ferguson (Eds.), *Crosscurrents in SLA and linguistic theory* (pp. 385-403). Amsterdam: Benjamins.

Warner, R. M. (2013). *Applied statistics: From bivariate through multivariate techniques* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Weigle, S.C. (2002). *Assessing writing.* Cambridge: Cambridge University Press.

Wendel, J. (1997). *Planning and second language narrative production* (Unpublished doctoral dissertation). Temple University, Japan.

Wickens, C. D. (1989). Attention and skilled performance. In Holding (Ed.), *Human skills* (pp. 71–105). New York: John Wiley.

Wigglesworth, G. (1997). An investigation of planning time and proficiency level on oral test discourse. *Language Testing, 14*(1), 85–106.

Wilkinson, A., Barney, G., Hanna, R, & Swan, M. (1980). *Assessing language development.* Oxford: Oxford University Press.

Winerip, M. (2005, May). SAT essay test rewards length and ignores errors. *The New York Times.* Retrieved from http://www.nytimes.com

Witte, S. P. (1987). Pre-text and composing. *College Composition and Communication, 38*, 397-425.

Wolfe-Quintero, K., Inagaki, S., & Kim, H.-Y. (1998). *Second language development in writing: Measures of fluency, accuracy, and complexity*. Honolulu, HI: University of Hawai'i, Second Language Teaching and Curriculum Center.

Yang, H. (2013). The case for being automatic: Introducing the automatic linear modeling (LINEAR) procedure in SPSS statistics. *Multiple Linear Regression Viewpoints, 39*(2), 27-37.

Yang, W. (2009). Topic effect on writing fluency and linguistic complexity of ESL writers and predictive values of writing fluency and linguistic complexity of ESL writers on writing scores (Unpublished course paper). Georgia State University, Atlanta, GA.

Yang, W., & Weigle, S. C. (2011, October). Lexical richness of ESL writing and the role of prompt. Paper presented at the 10th Conference for the American Association for Corpus Linguistics (AACL), Atlanta, GA.

Yang, W., Lu, X., & Weigle, S. C. (2012, March). Syntactic complexity of ESL writing, writing performance, and the role of topic. Paper presented at Georgetown University Round Table on Languages and Linguistics 2012 (GURT 2012), Washington, DC.

Yu, G. (2007). Lexical diversity in MELAB writing and speaking task performances. *Spaan Fellow Working Papers in Second or Foreign Language Assessment, 5*, 79-116.

Yu, G. (2010). Lexical diversity in writing and speaking task performances. *Applied Linguistics, 31*, 236-259.

Yun, Y. (2005). *Factors explaining EFL learners' performance in a timed essay writing test: A structural equation modeling approach* (Unpublished Doctoral dissertation). University of Illinois at Urbana-Champaign, Urbana, Illinois.

APPENDICES

Appendix A

A Pilot Validation Study of the Rhetorical Task Categories

In order to validate the rhetorical task categories of exposition, expo-argumentation, and argumentation used in the current study, a pilot validation study was conducted among 10 ESL professionals and graduate students in an Applied Linguistics graduate program which has two ESL teaching programs affiliated with it. The participants were first shown descriptions of three prompt categories, and after confirmation of clear understanding of the descriptions and the distinctions among the categories, they were given nine writing prompts and asked to sort the prompts into the three categories based on the category descriptions. In general, the participants found the descriptions clearly worded and self-explanatory. When there was anything unclear, the participants asked the researcher questions and gave her comments. To avoid association between terms and descriptions, “Category A”, “Category B”, and “Category C” were used rather than the actual terms of argumentation, expo-argumentation, and exposition. What follows are the descriptions used:

Category A:
The writing prompt mainly invites a writer to give personal opinions and judgment on a debatable issue or statement and to argue for a point a view on the issue/statement, based on facts, generalizations, reasoning, and/or inferences.

Category B:
The writing prompt mainly invites a writer to explain, to provide information about something, with personal opinions and judgment on the topic involved (but not to argue for a point of view on a debatable issue or statement), based on facts, generalizations, reasoning, and/or inferences.

Category C:
The writing prompt mainly invites a writer to explain and to provide information about something (not to give personal opinions or judgment or to argue on the topic), based on facts and generalizations of events and states.

Half way through the data collection, one of the participants suggested a wording choice for category A, so that categories A and B could be better distinguished, and the phrase of “to argue for a point a view” was changed to “to take a stand”, and the revised wording was used for the rest of the data collection. The following nine writing prompts, all on the same topic area of use of computers and the Internet, were presented to each participant with each prompt typed on a small piece of paper, and the participants read the prompts and matched each prompt with a prompt category whose description they had read by writing A, B, or C on the piece of paper for the prompt. Before the matching, the participants were informed that the writer population for the prompts would be Chinese university students in China who study English as an additional language. For the nine prompts, the first three are seen by the researcher as category A prompts (argumentative tasks), the second three as category B prompts (expo-argumentative tasks), and the third three as category C prompts (expository tasks).

Category A Prompts:

Computers and the Internet have improved the efficiency and quality of your learning as a university student. Do you agree or disagree with the statement?

Computers and the Internet have improved the efficiency and quality of learning for university students in your country. Do you agree or disagree with the statement?

Since computers and the Internet can promote information flow and human communication, they are bridging the gap in wealth between affluent areas and poor areas in your country. Do you agree or disagree with the statement?

Category B Prompts:

What do you think are the benefits and possible problems that computers and the Internet bring to you as a university student?

What do you think are the benefits and possible problems that computers and the Internet bring to university students in your country?

What do you think are the benefits and possible problems that computers and the Internet bring to people in poorer areas of your country where there is limited access to computers and the Internet?

Category C Prompts:

How do you use computers and the Internet in your life as a university student?

What are some ways that university students in your country use computers and the Internet?

What are some ways that people in poorer areas of your country use computers and the Internet?

For category B prompts, an earlier try-out of the validation on two other participants used “Discuss the benefits and possible problems …” rather than “What do you think are the benefits and possible problems …”; however, “Discuss …” prompt wording invited different interpretations since it was sometimes viewed as an invitation of showing factual knowledge that writers have learned from books, classes, or other reliable sources and other times viewed as an invitation of writers’ own construction of meaning on the topic. Since the writing tasks are independent writing tasks, with no expository reading materials to go along the tasks, it cannot be assumed that the writers have learned facts to answer the prompt questions. Further, since category B prompts are meant to be expo-argumentative tasks, writers’ opinions and judgment are inherent in the tasks. To avoid potential confusions, “What do you think are the benefits and possible problems …” were used instead of “Discuss the benefits and possible problems …” in the validation study.

With three prompt categories and three prompts for each category, there were a total of nine categorizations conducted by each participant. With 10 participants, there were 90 categorizations in total. The results of the categorizations were highly consistent, with 85 categorizations matching the ones the researcher had, achieving 94% consistency. Eight of the 10 participants had exactly the same categorizations for all prompts. The 6% inconsistent categorizations were from two participants only, with one having two categorizations that were different from the other participants’ and with the other having three categorizations that were different from the other participants’. The rather high consistency in the categorizations gives validity evidence for the prompt categories used in the current study. Further, some participants voluntarily labeled the categories: several called category C argumentative, and a couple of them named category C informational.

Appendix B

You can choose to complete the form in English **OR** Chinese.
您可以选择用英语**或者**中文填写下表。

Demographic Information Questionnaire (English)

Your name: ______________________ Your English teacher name: _____________________

Gender (check one): Female ___________ Male ____________

Age: ___________

Academic Status (check one):

Freshman _________ Sophomore _________ Junior _________ Senior _________

Major (or intended major): ________________________________

Number of years of learning English (in and out of school): _________ years ________ months

Have you ever stayed or studied in English-speaking countries? (check one): Yes ____ No ____
If yes, how long? _________ years ________ months ________ days

背景问卷 （中文）

姓名：______________________ 英语老师姓名：_____________________

性别 （选择一个）：女 ___________ 男 ___________

年龄：___________

学业状况：大学年级（选择一个）：大一_______ 大二_______ 大三_______ 大四_______

专业 （或意向的专业）：________________________________

学习英语的总共年数 （包括校内外的各种学习）：_________ 年_________ 月

你曾经在以英语为母语的国家待过或学习过吗？（选择一个）：有_____ 没有_____

如果有的话，多长时间：_________ 年_________ 月_________日

Appendix C

Writing Task 作文

Name: ________________Your English teacher's name: ________________ Date: __________

姓名：_____________________ 英语老师姓名：_________________ 日期：____________

DIRECTIONS

1. You will have 30 minutes to complete your essay.

2. Write an essay in response to the writing prompt provided with at least 150 words.

3. Your essay will be rated based on idea development and support in relation to the prompt and the task, organization and flow of ideas, and language use (in syntax, lexis, and etc.).

作文要求：

1. 请在 30 分钟内完成这篇作文。

2. 作文题目如下所列；作文长度要求：150 字以上。

3. 作文评分标准：与作文题和作文任务相关的写作内容的发展和支持、作文结构和语意连贯性、及语言的使用（句式、用词、语法等）。

Writing Prompt （作文题目）：

[Insert the writing prompt here.] [在这里提供作文题目。]

You can use the space below to do planning for your essay. Please write your essay from the next page. 您可以用下边的空白处计划您的写作。请从下页起开始写您的作文。

Appendix D

You can choose to complete the form in English **OR** Chinese. 您可以选择用英语**或者**中文填写下表。

Post-Writing Questionnaire (English)

1. You were given 30 minutes for the writing task. About how much time did it take you to complete your essay, including the time you used for checking and revising your essay?

   __________ minutes

2. Did you do planning before starting to write the essay? (check one): Yes ______ No ______

   If yes, about how much time did you spend on the planning? __________ minutes

   And how did you do the planning? ____

   ____

3. How interested were you in the writing topic you just wrote on? (Circle one)

   not interested at all ---- slightly interested ---- somewhat interested ---- greatly interested

4. For you, how difficult was the writing task you just completed? (Circle one) And indicate why.

   very easy ---- quite easy ---- a little difficult ---- very difficult

   Reasons for your choice: ____

   ____

5. How familiar are you with the rhetorical task of narration in English? (Circle one)

   not familiar at all ---- slightly familiar ---- somewhat familiar ---- very familiar

6. About how many hours do you use computers and the Internet per week? ___________ hours

7. Rank the difficulty of the writing task you just completed, together with three other tasks listed below. Consider how easy or how difficult it is for you to write on the topics in English.

   Use 1, 2, 3, and 4 to rank the difficulty, with 1 to indicate the easiest and 4 the most difficult.

   ______ What are some ways that university students in this country use computers and the Internet?

   ______ Computers and the Internet have improved the efficiency and quality of learning for university students in this country. Do you agree or disagree with the statement? Support your position with reasons.

   ______ Describe **one** of your experiences in which you used computers and/or the Internet for completing a course assignment or project or for studying for a school subject matter.

   ______ What do you think are the benefits and possible problems that computers and the Internet bring to university students in this country?

OR

Use 1, 2, and 3 to rank the difficulty, with 1 to indicate the easiest and 3 the most difficult.

______ What do you think are the benefits and possible problems that computers and the Internet bring to **you** as a university student?

______ What do you think are the benefits and possible problems that computers and the Internet bring to **people in underdeveloped areas of the world** where there is limited access to computers and the Internet?

______ What do you think are the benefits and possible problems that computers and the Internet bring to **university students** in this country?

[8. How familiar are you with underdeveloped areas in this country? (Circle one)

not familiar at all ---- slightly familiar ---- somewhat familiar ---- very familiar]

Notes:

1. For question 5, depending on which rhetorical task a writer got, that specific rhetorical task name (narration, exposition, exposition-argumentation, or argumentation) was used in the questionnaire the writer received.
2. For question 7, depending on which prompt a writer got, either the prompt set for rhetorical tasks or the prompt set for topic familiarity was used in the questionnaire the writer received; the writer rated the difficulty of the task he/she completed and the other tasks in the same cognitive complexity dimension. Half of the students who wrote on the shared task–benefits and problems of computers and the Internet for university students in China, received the rhetorical task prompt set, and the other half of them received the topic familiarity prompt set.
3. For question 8, only the students writing on the prompt of benefits and problems of computers and the Internet for people in underdeveloped areas of the world were presented with and asked to answer the question.

写作后问卷 （中文）

1. 您一共有 30 分钟完成这个写作任务。您实际花了大概多长时间写完您的作文的（包括您检查和修改作文的时间）？ __________ 分钟

2. 您在开始写这篇作文之前进行写作计划了吗？（选择一个）有______没有______

   如果有的话，您大概花了多长时间进行写作计划？__________ 分钟

   您是怎样进行写作计划的？____

   ____

3. 您对您刚刚完成的写作任务的话题有多大兴趣？（圈出您的选择）

   一点都没兴趣 ---- 有一点点兴趣 ---- 有一些兴趣 ---- 非常有兴趣

4. 对于您来说，您刚刚完成的写作任务是比较容易还是比较难？（圈出您的选择）并表明为什么。 非常容易 ---- 比较容易 ---- 有点难 ---- 非常难

   难度评价的原因：____

   ____

5. 您对英语记叙文熟悉吗？（圈出您的选择）

   一点都不熟悉 ---- 有一点熟悉 ---- 比较熟悉 ---- 非常熟悉

6. 您每个星期大概花多少个小时在使用电脑和网络上？___________ 小时

7. 请对下列四个写作任务（包括您刚刚完成的那个）排列他们的写作难度。请考虑如果您要用英文写这些作文的话，他们会有多容易或多难写。

   用 1，2，3，和 4 来排列他们的写作难度；1 表示最容易的，4 表示最难的。

______中国的大学生一般使用电脑和网络做些什么？

______电脑和网络提高了中国大学生学习的效率和质量。你同意还是不同意这个观点？用原因支持你的立场。

______描述**一次**你使用电脑和/或网络来完成某门课的作业、课题项目或用它们来学习有关学校某课程内容的经历。

______你认为电脑和网络给中国大学生带来了哪些益处和可能的问题？

或者

用 1， 2， 和 3 来排列他们的写作难度；1 表示最容易的，3 表示最难的。

______你认为电脑和网络给**你**（一个大学生），带来了哪些益处和可能的问题？
______你认为电脑和网络给**世界上比较贫困的、使用电脑和网络机会有限的地区的人民**带来了哪些益处和可能的问题？
______你认为电脑和网络给**中国大学生**带来了哪些益处和可能的问题？

[8. 您对中国的比较贫困的地区熟悉吗？（圈出您的选择）

一点都不熟悉 ---- 有一点熟悉 ---- 比较熟悉 ---- 非常熟悉]

Appendix E

Cloze Test 完型填空

姓名：______________________ 英语老师姓名：__________________ 日期：_____________

DIRECTIONS

1. Read the passage quickly to get the general meaning.
2. Write only one word in each blank next to the item number. Contractions are considered to be one word.
3. Check your answers.

You have 30 minutes to complete the cloze test.

完成这个完型填空的步骤：

1. 快速地阅览这篇文章，得知文章的大概意思。
2. 在每个题号边的空格处填写一个英语单词。缩写算一个单词。
3. 检查您的答案。

您有 30 分钟完成这个完型填空。

EXAMPLE （范例）: The boy walked up the street. He stepped on a piece of ice.

He fell (1) ***down***___ but he didn't hurt himself.

MAN AND HIS PROGRESS

Man is the only living creature that can make and use tools. He is the most teachable of living beings, earning the name of Homo sapiens. (1)____ ever restless brain has used the (2)____ and the wisdom of his ancestors (3)____ improve his way of life. Since (4)______ is able to walk and run (5)____his feet, his hands have always (6)____ free to carry and to use (7)_____ . Man's hands have served him well (8)_____ his life on earth. His development, (9)_____ can be divided into three major (10)____, is marked by several different ways (11)_____ life.

Up to 10,000 years ago, (12)____ human beings lived by hunting and (13)_____. They also picked berries and fruits, (14)_____ dug for various edible roots. Most (15)_____, the men were the hunters, and (16)_____ women acted as food gatherers. Since (17)_____ women were busy with the children, (18)_____ men handled the tools. In a (19)_____ hand, a dead branch became a (20)____ to knock down fruit or (21)____ for tasty roots. Sometimes, an animal (22)____ served as a club, and a (23)____ piece of stone, fitting comfortably into (24)____ hand, could be used to break (25)____ or to throw at an animal. (26)____ stone was chipped against another until (27)____ had a sharp edge. The primitive (28)____ who first thought of putting a (29)____ stone at the end of a (30)____ made a brilliant discovery: he (31)____ joined two things to make a (32)____ useful tool, the spear. Flint, found (33)____ many rocks, became a common cutting (34)_____ in the Paleolithic period of man's (35)____. Since no wood or bone tools (36)_____ survived, we know of this man (37)____ his stone implements, with which he (38)_______ kill animals, cut up the meat, (39)____ scrape the skins, as well as (40)____ pictures on the walls of the (41)_____ where he lived during the winter.

(42)____ the warmer seasons, man wandered on (43)____ steppes of Europe without a fixed (44)____, always foraging for food. Perhaps the (45)_____ carried nuts and berries in shells (46)____ skins or even in light, woven (47)_____ . Wherever they camped, the primitive people (48)_____ fires by striking flint for sparks (49)_____ using dried seeds, moss, and rotten (50)____ for tinder. With fires that he kindled himself, man could keep wild animals away and could cook those that he killed, as well as provide warmth and light for himself.

Answer keys

"Man and his progress" - answer keys

	Exact answer	Acceptable answer scoring would also include these possibilities
1	His	man's, our, the
2	Knowledge	accomplishments, culture, cunning, examples, experience(s), hands, ideas, information, ingenuity, instinct, intelligence, mistakes, nature, power, skill(s), talent, teaching, technique, thought, will, wit, words, work
3	to	
4	man	he
5	on	upon, using, with
6	been	felt, hung, remained
7	tools	adequately, carefully, conventionally, creatively, diligently, efficiently, freely, implements, objects, productively, readily, them, things, weapons
8	during	all, for, improving, in, through, throughout, with
9	which	also, basically, conveniently, easily, historically, however, often, since, that, thus
10	periods	areas, categories, divisions, eras, facets, groups, parts, phases, sections, stages, steps, topics, trends
11	of	for, in, through, towards
12	all	early, hungry, many, most, only, primitive, the, these
13	fishing	farming, foraging, gathering, killing, scavenging, scrounging, sleeping, trapping
14	and	often, ravenously, some, they
15	often	always, emphatically, important, nights, normally, of, times, trips
16	the	all, house, many, most, older, their, younger
17	the	all, many, married, most, often, older, primate, these
18	the	all, constructive, many, most, older, primate, tough, younger
19	man's	able, big, closed, coordinated, creative, deft, empty, free, human('s), hunter's, learned, needed, needy, person's, right, single, skilled, skillful, small, strong, trained
20	tool	club, device, instrument, pole, rod, spear, stick, weapon

21	dig	burrow, excavate, probe, search, test
22	bone	arm, easily, foot, head, hide, horn, leg, skull, tail, tusk
23	sharp	big, chipped, fashioned, flat, hard, heavy, large, rough, round, shaped, sizeable, small, smooth, soft, solid, strong, thin
24	the	a, his, man's, one('s)
25	nuts	apart, bark, bones, branches, coconuts, down, firewood, food, heads, ice, items, meat, objects, open, rocks, shells, sticks, stone, things, tinder, trees, wood
26	one	a, each, flat, flint, glass, hard, obsidian, shale, softer, some, the, then, this
27	it	each, one, they
28	man	being, creature, human, hunter, men, owner, people, person
29	sharp	glass, hard, jagged, large, lime, pointed, sharpened, small
30	stick	bone, branch, club, log, pole, rod, shaft
31	had	accidentally, cleverly, clumsily, conveniently, creatively, dexterously, double, easily, first, ingeniously, securely, simply, soon, suddenly, tastefully, then, tightly, would
32	very	bad, extremely, good, hunter's, incredibly, intelligent, long, modern, most, necessarily, new, portentously, quite, tremendously, useful
33	in	all, among, amongst, by, inside, on, that, using, within
34	tool	device, edge, implement, instrument, item, material, method, object, piece, practice, stone, utensil
35	development	age, ancestry, discoveries, era, evolution, existence, exploration, history, life, time
36	have	actually, apparently, ever
37	by	and, for, from, had, made, through, used, using
38	could	did, would
39	and	carefully, help, or, skillfully, then, would
40	draw	carve, create, drawing, engrave, hang, paint, painting, place, sketch, some, the
41	cave(s)	animals, place(s), room
42	in	and, during, with
43	the	across, aimless, all, barren, dry, flat, high, in, long, many, plain, stone, through, to,

		toward, unknown, various
44	home	appetite, camp, course, destination, destiny, diet, direction, domain, foundation, habitat, income, knowledge, location, lunch, map, meal, path, pattern, place, plan, route, supplement, supply, time, weapons
45	women	children, families, group, human, hunter, man, men, people, primitives, voyager, wanderers, woman
46	or	and, animal, animal's, covered, in, like, of, on, their, using, with
47	baskets	bags, blankets, chests, cloth(s), clothes, fabric, garments, hides, material, nets, pouches, sacks
48	made	began, built, lighted, lit, produced, started, used
49	and	also, by, occasionally, or, then, together, while
50	wood	bark, branches, dung, forage, grass, leaves, lumber, roots, skin, timber, tree(s)

Appendix F

Recruitment & Consent Procedures Explanations

Dear teachers,

Please say the following when announcing the research study in your class: (Please announce the research study in the class period preceding the one you have planned for the first data collection session.)

"WeiWei Yang is a PhD candidate in Applied Linguistics at Georgia State University in the U.S., and she invites you to participate in her dissertation research study. Her study examines the role of cognitive difficulty of essay writing tasks in second language writing quality and the language features of second language writing. WeiWei Yang also did her undergraduate study at … University. She sincerely wishes that you could take part in her study, but you do not have to be in the study. If you would like to take part, the research will involve a total of 1 hour and 10 minutes and will take place in two class periods of our English classes. The research will take place from our next class period. You will receive details of the research study in a consent form for the study at the beginning of our next class period. Based on the information presented in the consent form, you can decide whether you would like to take part in the study. If you do not wish to take part in the study, you can still do the study activities, but your data will not be collected; I will also prepare some study tasks related to our course for anyone who does not wish to work on the research activities."

Please say the following when giving the consent forms to your students at the beginning of the first class period you have planned for data collection:

"Please carefully read the consent form for the research study by WeiWei Yang. The consent form gives you the basic information about the study and your involvement in the study. If you agree to be in her study, please sign your name and put today's date in the space provided. You may take the timed writing task as an additional practice for your English essay writing. Both the timed writing task and the cloze test could be English-language practice for you. You do not have to be in the study. If you do not wish to be in the study, you do not need to sign on the consent form. If you do not wish to take part, you may still do the research activities, but your data will not be collected. If you do not wish to take part and also do not wish to work on the research activities, I have prepared some study tasks related to our course which you can work on while the research is taking place. For the students who agree to take part in WeiWei Yang's study, you will complete a cloze test in this class period. In the next class period [OR in our class on <weekday>], participating students will complete a background information sheet, a writing task, and a post-writing questionnaire. You may stop taking part at anytime."

Please say the following when giving out the research materials used in the second class period.

“If you have agreed to be in WeiWei Yang’s research by having signed the consent form in our last class [OR in our class on <weekday>] and you still wish to be in her study, you will complete in this class period the background information sheet, the writing task, and the post-writing questionnaire for the research. You do not have to be in the study, and you may stop taking part at anytime. If you have decided not to take part by having not signed the consent form, you may still work on the research activities, but your data will not be collected. If you have decided not to take part and also do not wish to do the research activities, I have prepared some study tasks related to our course which you can work on while the research is taking place.”

Thank you,
WeiWei Yang

招集研究参与者以及同意参与研究认可的程序

尊敬的各位老师,

请您在您班里宣布这个课题研究时向学生们说如下的内容：（请您在第一次课内数据收集之前的一节课内宣布这个课题研究。）

“杨微微是美国佐治亚州立大学应用语言学博士生；她竭诚邀请你参与她的博士论文研究。她的研究探索写作任务的认知难度对第二语言写作质量和作文中第二语言的使用的影响。杨微微也是在…大学读的本科。她真诚希望你能够参与她的研究，但是你并不必须参与这个研究。如果你愿意参与，这个研究总共需要 1 小时 10 分钟的时间，会在我们的英语课的两节课上进行。我们的下一节课上将开始这个课题研究。在下一节课开始时，你会收到一个参与研究的认可文件，上面提供了这个课题研究的具体细节。根据参与研究的认可文件上面的信息，你可以选择你是否想参与这个研究。如果你不希望参与这个研究，你仍然可以做这个研究的材料，但是你的资料时不会被收上来的；对于不愿意做研究材料的学生，我会准备一些与我们的课有关的学习任务让你上课做。”

请您在第一次课内数据收集的课开始时给学生发参与研究的认可文件，并向学生们说如下的内容：

“请认真阅读杨微微的课题研究的参与研究认可文件。这个认可文件提供了她的研究课题的相关信息以及有关你的参与的信息。如果你同意参与她的研究，请在文件上提供的空档处签名并写上今天的日期。你可以把那个限时写作任务看作一次额外的英语作文练习。那个限时写作任务和那个完型填空都可以给你提供一些英语语言练习。你并不必须参与这个研究。如果你不希望参与，那就不需要在参与认可文件上签名。如果你不希望参与这个研究，你仍然可以做这个研究的材料，但是你的资料时不会被收上来的。如果你不希望参与这个研究也不愿意做研究材料，我准备了一些与我们的课有关的学习任务让你做，当研究在课上进行时。对于同意参与杨微微的课题研究的学生，这节课上你将完成一个完型填空。在下一节课上[或者在周〈〉的课上]，参与这个研究的学生将完成一个学生背景问卷、一个写作任务以及一份写作后问卷。你可以在任何时间停止参与这个研究。”

请您在第二次课内数据收集的课上发研究材料时向学生们说如下的内容：

“如果你已经同意了参与杨微微的课题研究并在上节课上[或者在周〈〉的课上]签了参与研究的认可文件，而且你希望继续参与她的研究，你将在这节课上完成学生背景问卷、写作任务以及写作后问卷。你并不必须参与这个研究，你也可以在任何时间停止参与研究。如果你不希望参与这个研究，你仍然可以做这个研究的材料，但是你的资料时不会被收上来的。 如果你不希望参与这个研究也不愿意做研究材料，我准备了一些与我们的课有关的学习任务让你做，当研究在课上进行时。”

非常感谢您！
杨微微

Appendix G

Instructions to Teachers Recruiting Participants and Administering the Research Materials

Dear English Instructor,

I greatly appreciate your willingness to help me collect the data for my dissertation research in your English class! The data collection will involve class time in two of your class periods in Spring 2012 semester.

Data Collection in Class Period One (approximately 35 min. in total)

1. Consent Form (5 min.)
2. Cloze Test (30 min.)

Data collection procedures:

1) Distribute consent forms to your students. The students first read the consent form and then decide whether they would like to participate in the study. Those who decide to participate sign their names and write down that day's date in the provided space on the second page of the consent form.
2) Collect the consent forms, and give out the cloze tests to the students who have decided to participate. The cloze test is a 30-minute test. Please time the students.
3) When the 30 minutes is up for the cloze test, collect the tests from the participating students.

Notes:

1) The consent form is in Chinese. The directions for the cloze test are in both Chinese and English.
2) The cloze test is a validated one. In this research, it is used as a measure of the students' English proficiency level.
3) The students who decide not to participate in the study can also do the cloze test, but please do not collect their data. You may also arrange these students to do some other study tasks related to your course.
4) The cloze test is an interesting one. You may choose to let the students discuss their answers in the next class, and then you share the answer keys with your class. If you need to give out the collected cloze tests to the students, please make sure that they do not change their original answers. After the activity, please collect back the cloze tests again.

Data Collection in Class Period Two (approximately 40 min. in total):

1. Demographic Information Questionnaire (3 min.)
2. Writing Task (30 min.)
3. Post-Writing Questionnaire (5 min.)

Data collection procedures:

1) The demographic information questionnaire and the writing task materials are stapled together. Distribute the materials to the participating students.

2) The students first fill out the demographic information questionnaire; and then please require them to turn to the next page for the writing task at the same time. The writing task is a 30-minute one. Please time the students.
3) When the 30 minutes is up for the writing task, please collect the demographic information questionnaires and the writing task materials from the students.
4) After the writing task materials are collected, distribute the post-writing questionnaire to the participating students. The students complete the questionnaire in five minutes. Collect the questionnaires when finished.

Notes:

1) All the data collection materials for this second session are in both Chinese and English. The students can choose to read in either language and to respond to the questionnaires in either Chinese or English.
2) The participating students in your class will receive six different writing topics all related to use of computers and the Internet. You may choose to let your students share what they wrote in the next class. Please make sure that the students do not make changes to their original writing.
3) The students who decide not to participate in the study can also do the writing task and the post-writing questionnaire, but please do not collect their data. You may also arrange these students to do some other study tasks related to your course.

I greatly appreciate your help and support! If there is anything I can be of assistance, please contact me.

WeiWei Yang

PhD Candidate, Applied Linguistics

Georgia State University, USA

weiweiyang1@gmail.com

尊敬的各位老师，

非常感谢您帮助我在您的英语课上收集我做博士论文需要的数据！此次数据收集需要您 2012 春季学期两次课堂内时间的使用。

第一次课内数据收集 （总共约 35 分钟）：

1. 同意参与研究的认可（5 分钟）
2. 完型填空（30 分钟）

数据收集程序：

1) 把同意参与研究的认可的文件发给每个学生。学生们先阅读这个文件，然后决定他们是否要参与。同意参与的学生在文件第二页的空档处签名并写上当日的日期。
2) 把同意参与研究的认可文件收上来；同时把完型填空发给同意参与此研究的学生。这个完型填空限时 30 分钟。请计时。
3) 30 分钟的完型填空时间结束后，请把材料收上来。

注释：

1) 同意参与研究的文件是中文的。完型填空的要求是中英文双语的。
2) 这个完型填空是一个被考证过的测试。在这个研究中，它是用来当作学生英语水平的一个测试。
3) 不自愿参与这个研究的学生，他们也可以做这个完型填空；但是他们的数据不需要收上来。您也可以安排他们在课上时做一些其他的、与您课有关的学习任务。
4) 这个完型填空比较有意思。您可以选择在您的下一节课上让学生们讨论他们的答案，并把可能的答案和学生们分享。如果需要把收上来的完型填空发给学生，请确保学生们不改他们原先的答案。活动结束后，请再把完型填空收上来。

第二次课内数据收集 （总共约 40 分钟）：

1. 学生背景问卷 （3 分钟）
2. 写作任务 （30 分钟）
3. 写作后问卷 （5 分钟）

数据收集程序：

1) 学生背景问卷和写作任务会装订在一起；请同时发给学生。
2) 学生先完成背景问卷；然后要求学生们同时翻到第二张纸并开始写作任务。写作任务是 30 分钟。请计时。
3) 30 分钟的写作时间结束后，请把学生背景问卷和作文收上来。
4) 作文收上来后，把写作后问卷发给学生。学生们在 5 分钟内完成写作后问卷。学生做完后，把问卷收上来。

注释：

1) 所有的数据收集材料都是中英文双语的。学生们可以选择读中文或英文、用中文或英文完成问卷。
2) 您的班里的学生会收到六个不同的有关电脑和网络使用的写作题目。您可以选择在您的下一节课上让学生们交

流他们的写作内容。请确保学生们不改他们原先的写作。

3) 不自愿参与这个研究的学生，他们也可以做这个写作任务以及写作后问卷；但是他们的数据不需要收上来。您也可以安排他们做一些其他的、与您课有关的学习任务。

非常感谢您的帮助和支持！有什么我可以协助的地方，请跟我联系。

杨微微
博士生，应用语言学
美国佐治亚州立大学
weiweiyang1@gmail.com

Appendix H

TOEFL iBT Test Independent Writing Rubrics (Scoring Standards)

Score and Task description

Half-point ratings (e.g., 2.5) are given when an essay's quality falls in between the descriptors for two adjacent whole points.

Score 5 - An essay at this level largely accomplishes all of the following:

- effectively addresses the topic and task
- is well organized and well developed, using clearly appropriate explanations, exemplifications, and/or details
- displays unity, progression, and coherence
- displays consistent facility in the use of language, demonstrating syntactic variety, appropriate word choice, and idiomaticity, though it may have minor lexical or grammatical errors

Score 4 - An essay at this level largely accomplishes all of the following:

- addresses the topic and task well, though some points may not be fully elaborated
- is generally well organized and well developed, using appropriate and sufficient explanations, exemplifications, and/or details
- displays unity, progression, and coherence, though it may contain occasional redundancy, digression, or unclear connections
- displays facility in the use of language, demonstrating syntactic variety and range of vocabulary, though it will probably have occasional noticeable minor errors in structure, word form, or use of idiomatic language that do not interfere with meaning

Score 3 - An essay at this level is marked by one or more of the following:

- addresses the topic and task using somewhat developed explanations, exemplifications, and/or details
- displays unity, progression, and coherence, though connection of ideas may be occasionally obscured
- may demonstrate inconsistent facility in sentence formation and word choice that may result in lack of clarity and occasionally obscure meaning
- may display accurate but limited range of syntactic structures and vocabulary

Score 2 - An essay at this level may reveal one or more of the following weaknesses:

- limited development in response to the topic and task
- inadequate organization or connection of ideas
- inappropriate or insufficient exemplifications, explanations, or details to support or illustrate generalizations in response to the task
- a noticeably inappropriate choice of words or word forms

- an accumulation of errors in sentence structure and/or usage

Score 1 - An essay at this level is seriously flawed by <u>one or more of the following weaknesses:</u>

- serious disorganization or underdevelopment
- little or no detail, or irrelevant specifics, or questionable responsiveness to the task
- serious and frequent errors in sentence structure or usage

Score 0 - An essay at this level merely copies words from the topic, rejects the topic, or is otherwise not connected to the topic, is written in a foreign language, consists of keystroke characters, or is blank.

APPENDIX I

Writing Task-fulfillment Rating Rubrics

Narrative task:
Describe **one** of your experiences in which you used computers and/or the Internet for completing a course assignment or project or for studying for a school subject matter.

Task Fulfillment Rating:

code	**description**
1	The writer mainly described **one** of his/her experiences in which he/she used computers and/or the Internet for completing a course assignment or project or for studying for a school subject matter.
2	The writer made generalizations of his/her experiences in using computers and/or the Internet for completing course assignments or projects or for studying for a school subject matter, rather than describe one such experience.
3	The writer's approach does not belong to any of the above categories. Specify the writer's approach on the rating sheet.

Expository task:
What are some ways that university students in this country use computers and the Internet?

Task Fulfillment Rating:

code	**description**
1	The writer mainly described and made generalizations of ways that university students use computers and the Internet, although occasional judgment of the uses as being positive and/or negative might be involved.
2	The writer mainly approached the task by grouping and labeling the ways that university students use computers and the Internet as being positive and/or negative and providing a discussion of the benefits and/or problems of computer and Internet use by the university students.
3	The writer mainly described and made generalizations of ways that he/she himself/herself uses computers and the Internet, although occasional judgment of the uses as being positive and/or negative might be involved.
4	The writer's approach does not belong to any of the above categories. Specify the writer's approach on the rating sheet.

Expo-argumentative/ Impersonal-familiar task:
What do you think are the benefits and possible problems that computers and the Internet bring to university students in this country?

Task Fulfillment Rating:

code	description
1	The writer primarily approached the task collectively (from the perspectives of the 1st person plural – we, us, our, ours, and/or "university/college students" in general or by subgroup and/or the 3rd person plural – they, them, their, theirs) and discussed the benefits and problems that computers and the Internet brought to them and/or university/college students.
2	The writer primarily approached the task personally (from the perspectives of the 1st person singular – I, me, my, mine) and discussed the benefits and problems that computers and the Internet brought to him/her.
3	The writer approached the task collectively (from the perspectives of the 1st person plural – we, us, our, ours, and/or "university/college students" in general or by subgroup and/or the 3rd person plural – they, them, their, theirs) and personally (from the perspectives of the 1st person singular – I, me, my, mine) in a mixed and balanced manner and discussed the benefits and problems that computers and the Internet brought to him/her and to them or university/college students.
4	The writer primarily approached the task collectively (from the perspectives of people in general) and discussed the benefits and problems that computers and the Internet brought to people.
5	The writer's approach does not belong to any of the above categories. Specify the writer's approach on the rating sheet.

Note: primarily = 75% and up; mixed and balanced = 50%

Argumentative task:
Computers and the Internet have improved the efficiency and quality of learning for university students in this country. Do you agree or disagree with the statement? Support your position with reasons.

Task Fulfillment Rating:

code	description
1	The writer clearly stated his/her position on the debatable statement by choosing a side and wrote his/her support for the position.
2	The writer stated his/her position on the debatable statement by providing conditions for the truth value of the statement and wrote his/her support for the position.
3	The writer did NOT state his/her position on the debatable statement and only discussed the benefits and problems that computers and the Internet bring to university students in their learning.
4	The writer's approach does not belong to any of the above categories. Specify the writer's approach on the rating sheet.

Personal-familiar task:
What do you think are the benefits and possible problems that computers and the Internet bring to **you** as a university student?

Task Fulfillment Rating:

code	**description**
1	The writer primarily approached the task personally (from the perspectives of the 1st person singular – I, me, my, mine) and discussed the benefits and problems that computers and the Internet brought to him/her.
2	The writer primarily approached the task collectively (from the perspectives of the 1st person plural – we, us, our, ours, and/or "university/college students" in general or by subgroup and/or the 3rd person plural – they, them, their, theirs) and discussed the benefits and problems that computers and the Internet brought to them or university/college students.
3	The writer approached the task personally (from the perspectives of the 1st person singular – I, me, my, mine) and collectively (from the perspectives of the 1st person plural – we, us, our, ours, and/or "university/college students" in general or by subgroup and/or the 3rd person plural – they, them, their, theirs) in a mixed and balanced manner and discussed the benefits and problems that computers and the Internet brought to him/her and to them or university/college students.
4	The writer primarily approached the task collectively (from the perspectives of people in general) and discussed the benefits and problems that computers and the Internet brought to people.
5	The writer's approach does not belong to any of the above categories. Specify the writer's approach on the rating sheet.

Note: primarily = 75% and up; mixed and balanced = 50%

Impersonal-less familiar task:
What do you think are the benefits and possible problems that computers and the Internet bring to people in underdeveloped areas of the world where there is limited access to computers and the Internet?

Task Fulfillment Rating:

code	description
1	There was a lack of evidence that the writer considered the issue from the perspectives of the people in the underdeveloped areas. There was in general no indication that the writer's treatment of the issue was context-specific, with **basically no explicit** verbal contextualizations for the benefits and/or problems of computers and the Internet for the context specified, except for only mentioning "people in underdeveloped areas" once or twice. The writer's discussion revolving benefits and problems of computers and the Internet was very general.
2	The writer considered and addressed the issue from the perspectives of the people in the underdeveloped areas, in limited ways. There was some indication that the writer's treatment of the issue was context-specific, with **some but limited, explicit** verbal contextualizations for the benefits and problems of computers and the Internet for the context specified. But at other times, the writer's discussion revolving benefits and/or problems of computers and the Internet was rather general.
3	The writer considered and adequately addressed the issue from the perspectives of the people in the underdeveloped areas. There was clear and adequate indication that the writer's treatment of the issue was primarily context-specific, with **adequate and explicit** verbal contextualizations for **both** the benefits and problems of computers and the Internet for the context specified.
4	The writer's approach does not belong to any of the above categories. Specify the writer's approach on the rating sheet.

APPENDIX J

Table J1 *Accuracy, Fluency, Lexical Complexity, and Syntactic Complexity by Rhetorical Task and L2 Proficiency*

Construct/ Sub-construct	Measure	Narrative		Expository		Expo-Argu		Argumentative	
		L[a] (26)[b]	H[a] (35)[b]	L (32)	H (30)	L (35)	H (26)	L (34)	H (29)
Accuracy	errors per 100 words	3.53 (2.07)	2.23 (1.23)	3.77 (2.14)	2.53 (1.57)	3.61 (2.01)	2.32 (1.30)	3.08 (1.42)	1.77 (0.97)
Fluency	number of words per essay	176.23 (49.27)	215.66 (51.20)	174.72 (46.06)	236.67 (63.91)	203.83 (54.75)	251.65 (48.83)	195.68 (51.93)	235.55 (38.28)
Lexical diversity	vocd D	67.79 (18.58)	72.52 (17.53)	64.95 (13.94)	73.40 (18.21)	66.58 (13.03)	82.81 (19.59)	62.79 (16.78)	71.14 (18.15)
Lexical sophistication	proportion of sophisticated word types	0.09 (0.02)	0.12 (0.04)	0.10 (0.04)	0.14 (0.04)	0.10 (0.05)	0.15 (0.04)	0.09 (0.03)	0.13 (0.06)
Lexical density	lexical words/ all words	0.50 (0.03)	0.49 (0.04)	0.54 (0.03)	0.54 (0.04)	0.52 (0.03)	0.51 (0.03)	0.51 (0.03)	0.51 (0.03)
Overall sentence complexity	mean length of sentence	13.64 (2.11)	15.05 (3.49)	14.66 (3.16)	15.74 (2.24)	14.12 (2.81)	16.02 (2.94)	16.20 (4.33)	18.50 (3.51)
Overall T-unit complexity	mean length of T-unit	12.02 (1.61)	13.91 (2.75)	13.57 (2.43)	14.57 (2.36)	12.90 (2.09)	15.14 (2.84)	13.74 (2.05)	17.13 (2.79)
Clausal coordination	T-units per sentence	1.14 (0.12)	1.08 (0.12)	1.08 (0.12)	1.09 (0.18)	1.10 (0.15)	1.06 (0.07)	1.18 (0.25)	1.08 (0.09)
Finite subordination	dependent clauses per T-unit	0.42 (0.21)	0.43 (0.22)	0.39 (0.22)	0.38 (0.16)	0.35 (0.18)	0.51 (0.27)	0.42 (0.18)	0.52 (0.29)
Overall clause complexity	mean length of clause	8.52 (1.80)	9.88 (1.89)	9.89 (1.97)	10.67 (2.14)	9.34 (1.60)	10.12 (1.27)	9.57 (1.09)	11.44 (1.76)
Non-finite subordination	nonfinite elements per clause	0.39 (0.21)	0.40 (0.18)	0.38 (0.14)	0.44 (0.16)	0.27 (0.16)	0.31 (0.13)	0.36 (0.13)	0.48 (0.21)
Phrasal coordination	coordinate phrases per verb phrase	0.13 (0.08)	0.20 (0.11)	0.28 (0.16)	0.32 (0.14)	0.30 (0.13)	0.32 (0.16)	0.33 (0.14)	0.36 (0.11)
Noun-phrase complexity	complex NP per verb phrase	0.43 (0.19)	0.59 (0.22)	0.54 (0.18)	0.63 (0.22)	0.52 (0.24)	0.68 (0.18)	0.47 (0.13)	0.66 (0.22)

[a] L = lower proficiency; H = higher proficiency. [b] Enclosed in the parentheses are sample sizes.

APPENDIX K

Correlation Matrixes for Each Task

Table K1 *Correlation Matrix for the Narrative Task*

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. writing quality scores	1.00													
2. errors per 100 words	-0.25	1.00												
3. number of words per essay	0.60**	-0.20	1.00											
4. vocd D	0.36**	-0.08	0.15	1.00										
5. proportion of sophisticated word types	0.37**	-0.22	0.33**	0.31*	1.00									
6. lexical words/ all words	0.00	0.29*	-0.19	0.37**	0.05	1.00								
7. mean length of sentence	0.36**	-0.29*	0.30*	0.04	0.17	-0.16	1.00							
8. mean length of T-unit	0.45**	-0.27*	0.36**	0.18	0.27*	-0.09	0.83**	1.00						
9. T-units per sentence	-0.10	-0.11	-0.05	-0.23	-0.13	-0.13	0.44**	-0.12	1.00					
10. dependent clauses per T-unit	0.20	-0.25	0.29*	-0.16	-0.28*	-0.40**	0.38**	0.30*	0.19	1.00				
11. mean length of clause	0.21	-0.09	0.09	0.21	0.45**	0.17	0.47**	0.66**	-0.22	-0.44**	1.00			
12. nonfinite elements per clause	-0.16	-0.17	-0.05	-0.18	0.08	0.04	0.22	0.30*	-0.09	-0.22	0.57**	1.00		
13. coordinate phrases per verb phrase	0.23	-0.07	0.05	0.05	0.44**	0.13	0.43**	0.45**	0.04	-0.24	0.56**	0.04	1.00	
14. complex NP per verb phrase	0.33**	0.04	0.07	0.36**	0.38**	0.24	0.36**	0.52**	-0.20	-0.39**	0.74**	0.06	0.53**	1.00

** $p < 0.01$, 2-tailed; * $p < 0.05$, 2-tailed

Table K2 *Correlation Matrix for the Expository Task*

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. writing quality scores	1.00													
2. errors per 100 words	-0.44**	1.00												
3. number of words per essay	0.70**	-0.29*	1.00											
4. vocd D	0.21	-0.19	0.33*	1.00										
5. proportion of sophisticated word types	0.51**	-0.20	0.51**	0.38**	1.00									
6. lexical words/ all words	-0.20	0.19	-0.15	0.04	0.14	1.00								
7. mean length of sentence	0.25*	-0.21	0.15	-0.03	0.19	0.04	1.00							
8. mean length of T-unit	0.29*	-0.22	0.17	-0.03	0.31*	0.07	0.77**	1.00						
9. T-units per sentence	-0.03	0.00	-0.04	-0.07	-0.18	0.03	0.38**	-0.28*	1.00					
10. dependent clauses per T-unit	0.12	0.03	0.19	0.17	0.00	-0.27*	0.21	0.22	-0.06	1.00				
11. mean length of clause	0.31*	-0.25	0.11	-0.11	0.30*	0.15	0.39**	0.60**	-0.24	-0.51**	1.00			
12. nonfinite elements per clause	0.14	-0.07	0.00	-0.11	0.16	0.25	0.07	0.25	-0.25*	-0.35**	0.49**	1.00		
13. coordinate phrases per verb phrase	0.17	-0.21	0.05	-0.29*	0.10	0.13	0.27*	0.44**	-0.15	-0.25	0.64**	0.06	1.00	
14. complex NP per verb phrase	0.36*	-0.20	0.22	0.11	0.25*	0.05	0.28*	0.39**	-0.09	-0.35**	0.73**	0.01	0.50**	1.00

** p < 0.01, 2-tailed; * p < 0.05, 2-tailed

Table K3 *Correlation Matrix for the Expo-Argumentative, Impersonal Familiar Task*

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. writing quality scores	1.00													
2. errors per 100 words	-0.35**	1.00												
3. number of words per essay	0.62**	-0.24	1.00											
4. vocd D	0.38**	-0.17	0.53**	1.00										
5. proportion of sophisticated word types	0.55**	-0.25	0.52**	0.38**	1.00									
6. lexical words/ all words	-0.09	0.25*	-0.31*	0.00	-0.06	1.00								
7. mean length of sentence	0.40**	-0.11	0.39**	0.26*	0.30*	-0.23	1.00							
8. mean length of T-unit	0.49**	-0.11	0.43**	0.43**	0.41**	-0.10	0.86**	1.00						
9. T-units per sentence	-0.10	-0.02	-0.01	-0.28*	-0.18	-0.25	0.37**	-0.15	1.00					
10. dependent clauses per T-unit	0.30*	-0.07	0.43**	0.55**	0.26*	-0.15	0.61**	0.67**	-0.04	1.00				
11. mean length of clause	0.42**	-0.12	0.17	-0.04	0.33**	0.05	0.38**	0.52**	-0.21	-0.19	1.00			
12. nonfinite elements per clause	0.27*	-0.01	0.20	0.28*	0.16	-0.18	0.11	0.25	-0.26*	-0.10	0.52**	1.00		
13. coordinate phrases per verb phrase	0.10	-0.08	0.00	-0.28*	0.22	0.15	0.32*	0.34**	-0.01	-0.05	0.48**	-0.20	1.00	
14. complex NP per verb phrase	0.46**	-0.18	0.14	0.01	0.26*	0.17	0.26*	0.36**	-0.12	-0.04	0.65**	-0.06	0.39**	1.00

** *p* < 0.01, 2-tailed; * *p* < 0.05, 2-tailed

Table K4 *Correlation Matrix for the Argumentative Task*

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. writing quality scores	1.00													
2. errors per 100 words	-0.43**	1.00												
3. number of words per essay	0.53**	-0.28*	1.00											
4. vocd D	0.20	-0.06	0.22	1.00										
5. proportion of sophisticated word types	0.56**	-0.24	0.34**	0.25	1.00									
6. lexical words/ all words	0.23	0.06	-0.09	-0.13	0.19	1.00								
7. mean length of sentence	0.25*	-0.24	0.29*	0.01	0.25*	-0.16	1.00							
8. mean length of T-unit	0.41**	-0.47**	0.40**	0.04	0.34**	-0.13	0.70**	1.00						
9. T-units per sentence	-0.11	0.21	-0.06	-0.05	-0.03	-0.09	0.62**	-0.12	1.00					
10. dependent clauses per T-unit	0.13	-0.19	0.30*	0.20	0.18	-0.43**	0.45**	0.62**	-0.04	1.00				
11. mean length of clause	0.39**	-0.43**	0.13	-0.12	0.25	0.31*	0.37**	0.56**	-0.11	-0.25*	1.00			
12. nonfinite elements per clause	0.24	-0.39**	0.15	0.06	0.26*	0.20	0.21	0.32*	-0.06	-0.07	0.59**	1.00		
13. coordinate phrases per verb phrase	0.04	-0.06	0.07	-0.37**	-0.09	0.37**	0.15	0.31*	-0.13	-0.16	0.46**	0.01	1.00	
14. complex NP per verb phrase	0.42**	-0.28*	0.09	-0.11	0.31*	0.10	0.25*	0.43**	-0.14	-0.07	0.59**	-0.01	0.17	1.00

** $p < 0.01$, 2-tailed; * $p < 0.05$, 2-tailed

Table K5 *Correlation Matrix for the Personal Familiar Task*

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. writing quality scores	1.00													
2. errors per 100 words	-0.40**	1.00												
3. number of words per essay	0.63**	-0.27*	1.00											
4. vocd D	0.10	0.18	0.21	1.00										
5. proportion of sophisticated word types	0.42**	-0.29*	0.42**	0.15	1.00									
6. lexical words/ all words	-0.15	0.19	-0.18	0.21	0.17	1.00								
7. mean length of sentence	0.18	-0.17	0.19	0.04	0.25*	0.24	1.00							
8. mean length of T-unit	0.23	-0.15	0.22	0.06	0.34**	0.33**	0.93**	1.00						
9. T-units per sentence	-0.13	-0.05	-0.05	-0.07	-0.25*	-0.25*	0.20	-0.17	1.00					
10. dependent clauses per T-unit	0.06	0.07	0.20	0.26*	0.07	0.08	0.44**	0.50**	-0.15	1.00				
11. mean length of clause	0.28*	-0.25*	0.15	-0.11	0.35**	0.25*	0.58**	0.68**	-0.24	-0.18	1.00			
12. nonfinite elements per clause	0.06	-0.12	-0.07	0.08	0.21	0.34**	0.39**	0.48**	-0.26*	0.00	0.66**	1.00		
13. coordinate phrases per verb phrase	0.06	-0.19	0.04	-0.41**	0.25*	0.26*	0.40**	0.42**	0.01	-0.19	0.63**	0.10	1.00	
14. complex NP per verb phrase	0.37**	-0.19	0.39**	0.12	0.37**	0.23	0.46**	0.51**	-0.12	-0.08	0.65**	0.11	0.46**	1.00

** $p < 0.01$, 2-tailed; * $p < 0.05$, 2-tailed

Table K6 *Correlation Matrix for the Impersonal Less Familiar Task*

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. writing quality scores	1.00													
2. errors per 100 words	-0.32*	1.00												
3. number of words per essay	0.53**	-0.04	1.00											
4. vocd D	-0.09	0.01	0.00	1.00										
5. proportion of sophisticated word types	0.46**	-0.32*	0.33**	0.19	1.00									
6. lexical words/ all words	0.19	0.05	-0.03	0.21	0.08	1.00								
7. mean length of sentence	0.05	-0.17	0.09	0.03	0.17	-0.11	1.00							
8. mean length of T-unit	0.21	-0.25	0.18	0.15	0.28*	0.12	0.79**	1.00						
9. T-units per sentence	-0.17	0.02	-0.08	-0.12	-0.06	-0.31*	0.66**	0.06	1.00					
10. dependent clauses per T-unit	-0.12	-0.05	0.04	0.34**	0.06	-0.18	0.49**	0.53**	0.12	1.00				
11. mean length of clause	0.44**	-0.26*	0.19	-0.19	0.30*	0.38**	0.15	0.48**	-0.34**	-0.33**	1.00			
12. nonfinite elements per clause	0.11	-0.17	0.10	0.15	0.09	0.26*	-0.03	0.28*	-0.41**	-0.04	0.49**	1.00		
13. coordinate phrases per verb phrase	0.25*	-0.11	0.07	-0.53**	0.18	0.10	0.11	0.18	-0.05	-0.25*	0.53**	-0.13	1.00	
14. complex NP per verb phrase	0.45**	-0.13	0.17	-0.15	0.09	0.34**	0.24	0.40**	-0.08	-0.12	0.64**	-0.05	0.43**	1.00

** *p* < 0.01, 2-tailed; * *p* < 0.05, 2-tailed

APPENDIX L

Table L1 *Accuracy, Fluency, Lexical Complexity, and Syntactic Complexity by Topic Familiarity (On-task sample only)*

Construct/ Sub-construct	Measure	Personal-Familiar	Impersonal-Familiar	Impersonal-Less familiar	*F*	*p*	$\eta^2_{partial}$
Accuracy	errors per 100 words	2.80 (1.70)	3.01 (1.88)	2.93 (1.82)	0.10	0.90	0.00
Fluency	number of words per essay	218.57 (40.82)	222.51 (57.53)	237.11 (58.37)	1.01	0.37	0.02
Lexical diversity	vocd D	69.07 (14.09)	72.34 (17.06)	62.96 (21.25)	2.94	0.06	0.05
Lexical sophistication	proportion of sophisticated word types	0.12 (0.04)	0.12 (0.05)	0.09 (0.04)	4.42	0.01	0.07
Lexical density	lexical words/ all words	0.48 (0.03)	0.51 (0.03)	0.50 (0.04)	6.85	0.00*	0.11
Overall sentence complexity	mean length of sentence	14.65 (3.57)	14.87 (2.86)	17.10 (4.86)	4.63	0.01	0.08
Overall T-unit complexity	mean length of T-unit	13.55 (3.28)	13.80 (2.60)	15.32 (3.14)	3.67	0.03	0.06
Clausal coordination	T-units per sentence	1.08 (0.09)	1.08 (0.13)	1.11 (0.19)	0.56	0.57	0.01
Finite subordination	dependent clauses per T-unit	0.40 (0.20)	0.41 (0.23)	0.52 (0.27)	2.71	0.07	0.05
Overall clause complexity	mean length of clause	9.34 (2.34)	9.71 (1.52)	9.62 (1.72)	0.33	0.72	0.01
Non-finite subordination	nonfinite elements per clause	0.26 (0.19)	0.29 (0.15)	0.24 (0.18)	1.27	0.28	0.02
Phrasal coordination	coordinate phrases per verb phrase	0.27 (0.16)	0.31 (0.13)	0.31 (0.16)	0.74	0.48	0.01
Noun-phrase complexity	complex NP per verb phrase	0.54 (0.25)	0.59 (0.23)	0.75 (0.24)	6.90	0.00*	0.11

* *p* values are significant with Holm procedure adjustment, with overall α level set at 0.05.

APPENDIX M

Table M1 *Accuracy, Fluency, Lexical Complexity, and Syntactic Complexity by Topic Familiarity and L2 Proficiency*

Construct/ Sub-construct	Measure	Personal-Familiar		Impersonal-Familiar		Impersonal-Less familiar	
		L[a] (26)[b]	H[a] (35)[b]	L (35)	H (26)	L (30)	H (32)
Accuracy	errors per 100 words	3.23 (1.64)	2.11 (1.42)	3.61 (2.01)	2.32 (1.30)	3.53 (1.89)	2.24 (1.28)
Fluency	number of words per essay	201.03 (38.21)	238.14 (56.42)	203.83 (54.75)	251.65 (48.83)	208.83 (51.73)	228.97 (54.66)
Lexical diversity	vocd D	70.90 (16.82)	74.09 (18.91)	66.58 (13.03)	82.81 (19.59)	59.15 (17.01)	62.82 (20.19)
Lexical sophistication	proportion of sophisticated word types	0.11 (0.04)	0.14 (0.04)	0.10 (0.05)	0.15 (0.04)	0.07 (0.03)	0.11 (0.04)
Lexical density	lexical words/ all words	0.50 (0.03)	0.49 (0.04)	0.52 (0.03)	0.51 (0.03)	0.48 (0.04)	0.51 (0.03)
Overall sentence complexity	mean length of sentence	14.47 (3.10)	16.09 (3.10)	14.12 (2.81)	16.02 (2.94)	15.72 (5.39)	16.99 (2.93)
Overall T-unit complexity	mean length of T-unit	13.21 (2.61)	14.79 (2.97)	12.90 (2.09)	15.14 (2.84)	13.96 (2.98)	15.79 (2.79)
Clausal coordination	T-units per sentence	1.10 (0.09)	1.09 (0.08)	1.10 (0.15)	1.06 (0.07)	1.11 (0.20)	1.08 (0.10)
Finite subordination	dependent clauses per T-unit	0.33 (0.18)	0.45 (0.22)	0.35 (0.18)	0.51 (0.27)	0.45 (0.27)	0.47 (0.23)
Overall clause complexity	mean length of clause	9.53 (1.65)	10.07 (2.35)	9.34 (1.60)	10.12 (1.27)	9.10 (1.48)	10.32 (1.74)
Non-finite subordination	nonfinite elements per clause	0.27 (0.15)	0.29 (0.20)	0.27 (0.16)	0.31 (0.13)	0.19 (0.17)	0.23 (0.15)
Phrasal coordination	coordinate phrases per verb phrase	0.28 (0.14)	0.31 (0.18)	0.30 (0.13)	0.32 (0.16)	0.30 (0.15)	0.37 (0.16)
Noun-phrase complexity	complex NP per verb phrase	0.55 (0.18)	0.63 (0.22)	0.52 (0.24)	0.68 (0.18)	0.65 (0.22)	0.81 (0.26)

[a] L = lower proficiency; H = higher proficiency. [b] Enclosed in the parentheses are sample sizes.