June 2023

The Relationships Between L1, Writing Quality, and Complexity, Accuracy, and Fluency in L2 Writing

Tuc C. Chau
University of South Florida

The Relationships Between L1, Writing Quality, and Complexity, Accuracy, and Fluency in L2 Writing

by

Tuc C. Chau

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Department of World Languages
College of Arts and Sciences
University of South Florida

Major Professor: Matt Kessler, Ph.D.
Amanda Huensch, Ph.D.
Brandon Tullock, Ph.D.
Wei Zhu, Ph.D.

Date of Approval:
May 3, 2023

Keywords: writing proficiency, writing assessment, linguistic features, corpus linguistics

DEDICATION

To my Dad, who instilled in me the dream of studying in the U.S.

ACKNOWLEDGMENTS

I would not have come this far without the support of many people, near or far.

The completion of my dissertation is largely due to my major professor, Dr. Matt Kessler. His frequent contact helped me stay organized and focused, and his prompt feedback on my writing propelled me forward. I owe him a lot.

I would also like to thank my chair, Dr. Ippokratis Kantzios, and the other members of my committee, Dr. Amanda Huensch, Dr. Brandon Tullock, and Dr. Wei Zhu, for their thoughtful feedback on my work and support for my defense.

In particular, I would like to express my gratitude to Dr. Huensch, from whom I have learned so much. Without her continued guidance and support, I would not have been able to hone my skills and conduct rigorous research. I am blessed to have her as a mentor for the past years.

I am also grateful to Dr. Nicole Tracy-Ventura for guiding me in my early days at the University of South Florida and helping me lay the first bricks for my dissertation.

I would like to thank my colleagues and mentors at INTO for providing me with a supportive and collaborative environment. Their insights and expertise made me a better teacher.

I cannot forget my graduate peers, Jessica Giovanni, Mark Lane-Holbert, Oksana Bomba, Sean Farrell, and Shinji Shimoura, for sharing life moments and academic struggles with me. I also wish I could thank each of my friends who have been by my side throughout this journey, making it more memorable and enjoyable.

Finally, I owe it to my parents for raising and educating me. I love my family, mom, dad, and bro.

This project was supported by the University of South Florida Dissertation Completion Fellowship, so I would like to extend my appreciation to all that offered me this precious opportunity.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Abbreviations	Definitions
AI	Artificial intelligence
BNC	British National Corpus
CALF	Syntactic complexity, accuracy, lexical complexity, and fluency
CEFR	Common European Framework of Reference for Languages
CLAN	Computerized Language Analysis
COCA	Corpus of Contemporary America English
EAP	English for Academic Purposes
EFL	English as a Foreign Language
ESL	English as a Second Language
ETS	Educational Testing Service
ICLE	International Corpus of Learner English
L1	First language
L2	Second language
LCA	Lexical Complexity Analyzer
LOCNESS	Louvain Corpus of Native English Essays
NS	Native speakers
NNS	Non-native speakers
NLI	Native language identification
RQ	Research question
SCA	Syntactic Complexity Analyzer
SLA	Second language acquisition
TAALED	Tool for the Automated Analysis of Lexical Diversity
TAALES	Tool for the Automated Analysis of Lexical Sophistication
TAASSC	Tool for the Automatic Analysis of Syntactic Sophistication and Complexity
TOEFL	Test of English as a Foreign Language
VAC	Verb-argument construction

ABSTRACT

The purpose of the current dissertation is to map the relationships between first language (L1), writing quality, and syntactic complexity, accuracy, lexical complexity, and fluency (CALF) in second language (L2) writing. CALF are characteristics of language production that have been of significant interest in L2 writing research for the past few decades. Though they have been extensively studied as dependent variables that may vary as a function of other factors, they have been rarely studied together, much less in relation to L1 as an independent variable. Thus, this study explored the effects of L1 and writing quality, operationalized as score levels, on all four dimensions of CALF and the predictive power of CALF measures on writing quality. Adopting a quantitative, corpus-based approach, I collected 1,683 essays from the Educational Testing Service (ETS) Corpus of Non-Native Written English (TOEFL11) for analysis. The corpus is comprised of essays written by speakers of 11 non-English native languages (Arabic, Chinese, French, German, Hindi, Italian, Japanese, Korean, Spanish, Telugu, and Turkish) as part of an international test of academic English proficiency – TOEFL (Test of English as a Foreign Language). The selected essays were controlled for topic and collapsed into three score levels: low, medium, and high. They were automatically processed for 14 syntactic complexity measures, five lexical complexity measures, and one fluency measure using different automated tools. Approximately 20% (329 essays) were hand-coded for six accuracy measures. Statistical tests revealed that there were significant differences between L1s in most CALF measures in all score levels. Text length (W/Tx) was found to differentiate score levels in all L1s. Other relatively consistent indicators of score levels across L1s are the total number of errors and

lexical diversity measures such as the index of lexical diversity (D) and the measure of textual lexical diversity (MTLD). Multinominal regression models output mean length of sentence (MLS), the number of coordinate phrases per clause (CP/C), lexical density (LD), MTLD, lexical sophistication (LS1), and W/Tx as predictors of high-quality writing. Overall, results showed that CALF measures varied significantly across L1 backgrounds and score levels with several measures being predictive of the writing quality of a heterogenous group of L2 writers. These findings suggest that CALF should be examined together when assessing L2 writing and that L1 background is an important factor to consider when studying CALF in L2 writing. It is also necessary to tailor L2 instruction and assessment to address the unique challenges learners from different L1s face.

CHAPTER ONE: INTRODUCTION

Over the last few decades, the constructs of complexity, accuracy, and fluency or more specifically syntactic complexity, accuracy, lexical complexity, and fluency (CALF) have thrived as research variables in the fields of second language acquisition (SLA) and L2 writing. Researchers interested in exploring the effects of L2 instruction, task design, or individual differences on language production have included CALF as dependent variables in their studies (e.g., Bulté & Housen, 2014; Derwing & Rossiter, 2003; Ellis & Yuan, 2004; Yuan & Ellis, 2003). Complexity generally refers to the range and sophistication of grammatical structures and vocabulary that surface in language production, accuracy to the absence of errors, and fluency to the pace with which language is produced (Ortega, 2003; Polio, 2001; Wolfe-Quintero et al., 1998). CALF originally derived from the grammatical complexity and accuracy measures developed in L1 acquisition research to "expediently and reliably gauge proficiency in an L2" (Larsen-Freeman, 1978, p. 469). Nevertheless, it was not until the 1990s that CALF were brought together in a new proficiency model since their debut in SLA research in the 1970s. This model with CALF as its core constructs (Skehan, 1996, 1998) complements the traditional four skills model and sociolinguistic and cognitive models of L2 proficiency (Bachman, 1990; Bialystok, 1994; Canale & Swain, 1980).

CALF measures typically emerge as ratios, frequencies, and formulas (Norris & Ortega, 2009). Their use as indices of L2 proficiency and development has been justified both theoretically and empirically. In theory, CALF allow L2 proficiency to be measured in an objective, quantitative, and verifiable way. They can also effectively address the multifaceted

nature of L2 proficiency (Housen et al., 2012). They have been claimed to reflect different aspects of L2 proficiency such as the internalization of new L2 features (i.e., greater complexity), the modification of L2 knowledge (i.e., higher accuracy), and the consolidation and proceduralization of such knowledge (i.e., better fluency; De Graaff & Housen, 2009; Skehan, 1998, 2003). Empirically, CALF have been labelled as distinct and competing dimensions of L2 performance by factor analyses (Norris & Ortega, 2009; Ortega, 1995; Skehan & Foster, 1997, 2001), meaning for any claims about L2 learners' proficiency to be made, all the dimensions must be taken into account (Housen et al., 2012).

In the area of L2 writing, CALF, particularly syntactic complexity, have been mainly examined from four perspectives: language development, language performance, language proficiency, and writing quality (Barrot & Agdeppa, 2021). The one holding the most attention is language development as it is posited that syntactic complexity indexes the growth in learners' linguistic repertoire and their ability to utilize additional linguistic resources to communicate successfully (Ortega, 2015). Studies measuring CALF in relation to language development thus investigate whether CALF measures can validly and reliably capture L2 writing development over time (e.g., Bulté & Housen, 2014; Crossley & McNamara, 2014; Yoon & Polio, 2017). Similarly, CALF are viewed as indices of language proficiency, with higher CALF indices indicating higher proficiency. Studies adopting this view, however, do not necessarily examine the same group of language learners over time like studies of language development but different groups of learners across proficiency levels (e.g., Kuiken & Vedder, 2019; Lu, 2011; Martínez, 2018; Vo & Barrot, 2022). The third domain in which CALF are usually measured is language performance, with studies researching the variation of CALF measures based on cognitive factors involved in a writing task (i.e., task complexity; Amiryousefi, 2016; Johnson, 2017;

Kuiken & Vedder, 2007; Wigglesworth & Storch, 2009). Finally, as indices of writing quality, certain CALF measures are believed to be capable of distinguishing between poorly rated and highly rated essays (e.g., Casal & Lee, 2019; McNamara et al., 2010; Taguchi et al., 2013). In other words, higher graded papers are generally expected to demonstrate a higher amount of CALF. This last use of CALF measures is my focus in the current study.

Although the body of research on the relationship between CALF and L2 writing quality has built up recently, it is still lacking in the comprehensiveness of CALF constructs. Most studies either focus on syntactic complexity measures or examine only a selected number of CALF measures, which leads to a limited understanding of how different constructs of CALF differentiate writing quality in a sample. Accuracy and fluency should be included more in this line of research in addition to syntactic complexity, for example, to attain a more holistic view of and draw more definite conclusions about the correlation between CALF and L2 writing quality (Foster & Wigglesworth, 2016; Polio, 2017).

Aside from studying CALF constructs together, there have been suggestions to account for L1 as a moderating variable in L2 writing research (Lu & Ai, 2015; Ortega, 2015). For instance, Lu and Ai (2015, p. 26) concluded that "learners with different L1 backgrounds, even for those at the same or comparable proficiency levels, may not develop in the same ways in all areas" after measuring syntactic complexity in 1,400 argumentative essays written by college-level English as a Foreign Language (EFL) learners with seven different L1 backgrounds from the International Corpus of Learner English Version 2.0 (ICLE 2.0; Granger et al., 2009) and the Louvain Corpus of Native English Essays (LOCNESS; Granger, 1996). Another example is Murakami et al.'s (2013) investigation into cross-linguistic influence on accuracy. He analyzed 3,000 essays from the Cambridge Learner Corpus, which were sampled across seven L1 groups,

and found consistently higher accuracy levels for the L1s that mark a given morpheme compared with those that do not. Overall, the findings from these studies provide robust evidence that L1 influence cannot be left unchecked in the research designs of CALF studies in L2 writing (Ortega, 2015). However, such influence is mostly ignored in the current research atmosphere as studies tend to treat different L1 groups as one holistic non-native speaker (NNS) group or compare CALF differences between native speaker (NS) and NNS groups rather than between L1 groups. To fill the gaps in previous research, the current study examines all CALF constructs as indices of L2 writing quality while considering the potential influence of L1 on CALF measures. It specifically seeks to answer the following research questions:

1. To what extent do CALF measures vary across L1 backgrounds in each score level?
2. To what extent do CALF measures vary across score levels in each L1 group?
3. To what extent can CALF measures and L1 backgrounds predict score levels?

By mapping the relationships between L1, writing quality, and CALF in L2 writing, the current study contributes to a better understanding of CALF measures as indices of L2 writing quality. It will help teachers design L1-specific interventions. Researchers will also be more informed to decide whether L1 should be controlled for in future research.

In addition to this chapter that presents an overview of the study (Chapter One), this dissertation is organized into five other chapters. Chapter Two reviews the relevant literature and previous studies on CALF in L2 writing. Chapter Three explains the methodology of the study. Chapter Four reports the results of quantitative analyses, whereas Chapter Five discusses the findings. Finally, Chapter Six highlights the findings' implications for L2 writing assessment, pedagogy, and research. It also acknowledges the study's limitations and provides directions for future research.

CHAPTER TWO: LITERATURE REVIEW

This chapter is divided into three sections, corresponding to the three main foci of this dissertation. The first reviews literature regarding the linguistic constructs of CALF, and how they have been defined and operationalized in L2 writing research. The second reviews literature regarding the relationship between CALF and L2 writing quality. The last section reviews literature regarding the relationship between CALF and L2 writers' L1.

Defining and Operationalizing CALF in L2 Writing

Defining and operationalizing CALF have received significant attention over the years, as numerous researchers have put forward different measures. Nevertheless, the conceptualization of CALF as constructs is only one among many major challenges CALF researchers face, including the operationalization of CALF, the interrelationship among CALF components, the cognitive, linguistic, and psycholinguistic correlates of CALF, and the extrinsic factors affecting CALF (Housen & Kuiken, 2009; Housen et al., 2012).

True to its name, perhaps complexity is the most complex component of CALF (Housen & Kuiken, 2009; Pallotti, 2009). Pallotti (2009) gave an account of complexity by dissecting the word *complex*. According to him, it has three meanings. The first one is completely structural, which teachers and researchers use to distinguish simple and complex grammatical structures. The second meaning is what is perceived as *difficult* and *cognitively demanding* in CALF studies. Finally, *complex* is identified as "acquired late," meaning a complex structure may require many cognitive resources to produce and thus takes time to be internalized (Pallotti, 2009, p. 593).

Another approach to defining complexity is dividing it into task complexity and language, or more specifically, L2 complexity (Robinson, 2001). Task complexity refers to the processing demands of tasks, resulting from task structure and design together with learners' available resources. This type of complexity is preferably interpreted by some scholars as objective difficulty, indicating that the difficulty perceived is inherent to the task (Pallotti, 2009). Meanwhile, L2 complexity can be understood as cognitive complexity and linguistic complexity, both of which refer to properties of language features (items, patterns, structures, rules) or language subsystems (phonological, morphological, syntactic, lexical) (Housen & Kuiken, 2009). Cognitive complexity is based on learners' perceptions, and linguistic complexity is based on the language system. The former is a broader concept than the latter as it is determined by both subjective, learner-dependent factors such as aptitude, memory span, motivation, L1 background and more objective factors such as input saliency and linguistic complexity itself (Housen & Kuiken, 2009). In other words, linguistic complexity can affect cognitive complexity. It is "the size, elaborateness, richness, and diversity" of learners' L2 and at the same time, the structural complexity, including formal and functional complexity, of individual L2 features (Housen & Kuiken, 2009, p. 464). It can be categorized as grammatical complexity (syntactic complexity and morphological complexity) and lexical complexity (Bulté & Housen, 2012), of which syntactic complexity and lexical complexity are investigated in this dissertation as dependent variables.

Syntactic Complexity

When introducing her landmark synthesis of the relationship between syntactic complexity and L2 proficiency in college-level writing, Ortega (2003) encapsulates the definition, significance, and uses of syntactic complexity. In a broad sense, syntactic complexity involves

the range and degree of sophistication of syntactic structures demonstrated in language production. It is an important construct in L2 research due to the assumption that language development leads to, among other processes, the expansion of L2 learners' syntactic repertoire and their ability to use that repertoire properly for different purposes. In L2 writing research, syntactic complexity measures have been used to evaluate instructional effects on grammatical development and/or writing ability, to investigate task-related differences in L2 writing, and to assess variation in L2 texts produced by learners over time and across proficiency levels.

Syntactic complexity has increasingly been conceptualized as a multi-dimensional construct (Norris & Ortega, 2009), which can be interpreted on sentential, clausal, and phrasal levels (see Figure 1).

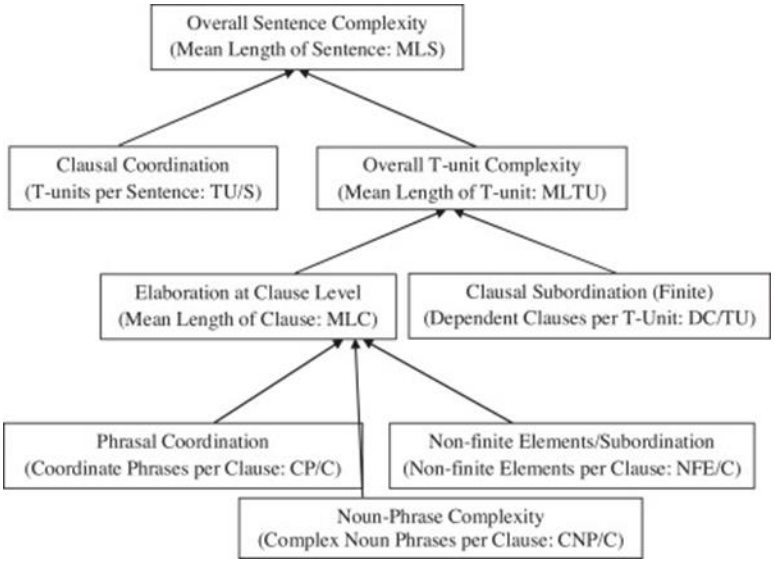

Figure 1. A Multi-Dimensional Representation of Syntactic Complexity (adopted from Yang et al., 2015).

Note. The measures in the parentheses are seen by Yang et al. (2015) as best operationalizations of the proposed syntactic complexity dimensions. MLS and MLT were labeled as global complexity measures and the other six measures as local-level complexity measures.

The representation above incorporated Norris and Ortega's (2009) suggestion that for syntactic complexity, L2 researchers should at least measure global or overall complexity, complexity by subordination, complexity by subclausal or phrasal elaboration, and possibly complexity by coordination. In fact, the conceptualization and operationalization of syntactic complexity are far from being consistent. Ortega (2003), for example, found that mean length of T-unit[1] (MLT) was the only measure shared by the six longitudinal L2 writing studies included in her synthesis. Across the other 21 cross-sectional studies she reviewed, MLS, MLT, MLC, mean number of T-units per sentence (T/S), mean number of clauses per T-unit (C/T), and mean number of dependent clauses per clause (DC/C) were the most popular measures of syntactic complexity. Critiquing the heavy reliance on T-units and clausal subordination, Biber et al. (2011) considered 28 different grammatical complexity features in academic writing against conversation, many of which deal with lexico-grammatical information at the word and phrase levels. They found that nearly all clausal subordination measures appeared more frequently in conversation than academic writing. They concluded that no single measure would satisfactorily capture complexity, and that measures other than clausal subordination and T-units must be developed to represent non-clausal features embedded in noun phrases – the most crucial types of complexity instruments in academic writing.

More recently, when synthesizing and meta-analyzing task-based L2 writing studies from 1998 to 2017, Johnson (2017) examined what CALF metrics were used. His examination of 20 studies yielded several notable results. First, complexity by subordination and global complexity were major concentrations of research with C/T and MLT being the most and second most reported metrics, respectively. Second, only three studies in the sample explored metrics on the

[1] T-unit is a term coined by Kellogg Hunt (1965), indicating a main clause and any dependent clause attached to it.

global, clausal, and phrasal complexity levels. Finally, only one study reported the range of forms produced as a metric. In general, Johnson's (2017) results together with previous findings showed that many L2 writing studies employed a small range of syntactic complexity measures that do not fully reflect the multidimensionality of syntactic complexity.

The current availability of automated tools for syntactic complexity analysis sheds additional light on the conceptualization of syntactic complexity. The fact that researchers have different approaches to defining and operationalizing syntactic complexity has prompted them to develop and/or use certain tools to analyze it. For instance, Coh-Metrix (Graesser et al., 2004; McNamara et al., 2014), in addition to measuring cohesion and coherence features of texts, can report on 15 syntactic complexity measures (i.e., words before main verb, number of modifiers per noun phrase, Minimal Edit Distance, sentence syntax similarity) and syntactic pattern density measures (i.e., phrase/agentless passive voice/negation/gerund/infinitive density), many of which have proven suitable for the task of investigating L2 writing syntactic complexity (e.g., see Crossley & McNamara, 2014). A more specialized tool is the L2 Syntactic Complexity Analyzer (SCA) with 14 measures representing length of production unit, amount of subordination, amount of coordination, phrasal sophistication, and overall sentence complexity (Lu, 2010), which match with the four dimensions of syntactic complexity proposed by Norris and Ortega (2009). These measures are claimed to either have a significant effect on L2 proficiency, as demonstrated by at least one previous study, or be recommended by Wolfe-Quintero et al. (1998) for further research (Ai & Lu, 2013; Lu, 2017).

Table 1

Syntactic Complexity Measures in SCA (adapted from Lu, 2017)

Category	Measure	Label
Length of production unit	Mean length of clause	MLC
	Mean length of sentence	MLS
	Mean length of T-unit	MLT
Amount of subordination	Number of clauses per T-unit	C/T
	Complex T-unit ratio	CT/T
	Number of dependent clauses per clause	DC/C
	Number of dependent clauses per T-unit	DC/T
Amount of coordination	Number of coordinate phrases per clause	CP/C
	Number of coordinate phrases per T-unit	CP/T
	Number of T-units per sentence	T/S
Degree of phrasal sophistication	Number of complex nominals per clause	CN/C
	Number of complex nominals per T-unit	CN/T
	Number of verb phrases per T-units	VP/T
Overall sentence complexity	Number of clauses per sentence	C/S

Among the 14 syntactic complexity measures (see Table 1), MLT, C/T, and DC/C are the most correlated with proficiency ($r > 0.65$; Lu, 2011) or best show an overall effect for proficiency with a significant difference between three or more adjacent proficiency levels ($p < 0.05$; Lu, 2011). More recently, Kyle and Crossley (2017) and Mostafa and Crossley (2020) combined the same 14 syntactic complexity measures with verb-argument construction (VAC) measures of syntactic sophistication that reflect usage-based perspectives of language learning. All of them can be computed using the Tool for the Automatic Analysis of Syntactic

Sophistication and Complexity (TAASSC; Kyle, 2016; Kyle & Crossley, 2017). Syntactic sophistication indices are calculated based on main verb lemmas (e.g., to give), VACs (e.g., subject-verb-indirect object-direct object), verb-VAC combinations (e.g., subject-to give-indirect object-direct object), and their frequencies in all written registers in the Corpus of Contemporary America English (COCA; Davies, 2008): fiction, magazines, newspapers, and academic texts (Mostafa & Crossley, 2020). A summary of these indices can be found in Table 2.

Table 2

Syntactic Sophistication Measures in TAASSC (adopted from Kyle & Crossley, 2017)

	Main verb lemma frequency	VAC frequency	Verb-VAC combination frequency
Mean token score	✓	✓	✓
Mean token score (log transformed)	✓	✓	✓
Standard deviation token score	✓	✓	✓
Standard deviation token score (log transformed)	✓	✓	✓
Mean type score	✓	✓	✓
Proportion of items attested in corpus	✓	✓	✓
Total	6	6	6

On the clausal and phrasal levels, researchers usually employ noun-related indices as measures of complexity. Ansarifar et al. (2018), for example, addressed the frequency and

distribution of 16 specific noun phrase features, which were adopted from Biber et al.'s (2011) developmental stages of syntactic complexity (see Table 3). Lan et al. (2019) similarly chose to include 11 noun modifiers in Biber et al.'s (2011) index in their analysis of Chinese first-year compositions. Both studies, however, excluded the representative feature of the first stage of development (probably due to the proficiency level of the participants), which is finite complement clauses (*that* and *wh*-) controlled by extremely common verbs (e.g., *think*, *know*, *say*). Occasionally, syntactic complexity measures are divided into large-grained and fine-grained measures (Jiang et al., 2019). Large-grained measures are the 14 SCA measures, while fine-grained measures refer to the grammatical structures related to subordinate clauses (adverbial clauses, complement clauses, and relative clauses) and noun modifiers (possessive modifiers, compound nouns, adjectival modifiers, prepositional phrases as attributes, and adjectival relative clauses).

Table 3

Hypothesized Developmental Stages for Complexity Features (adapted from Biber et al., 2011)

Stage	Grammatical structure(s)	Example
1	Finite complement clauses (*that* and *wh*-) controlled by extremely common verbs (e.g., think, know, say)	We never quite **know** what to make of him.
2	Finite complement clauses controlled by a wider set of verbs	I'd **forgotten** that he had just testified on that one.
	Finite adverbial clauses	I'm assuming I gained weight because things are a little tighter than they used to be.
	Nonfinite complement clauses, controlled by common verbs (especially want)	I don't **want** to fight with them about it.
	Phrasal embedding in the clause: adverbs as adverbials	He's so confused anyway.
	Simple phrasal embedding in the noun phrase: attributive adjectives	It certainly has a nice flavor.
3	Phrasal embedding in the clause: prepositional phrases as adverbials	He seems to have been hit on the head.
	Finite complement clauses controlled by adjectives	It seemed quite **clear** that no one was at home.
	Nonfinite complement clauses controlled by a wider set of verbs	The snow began to fall again.
	That relative clauses, especially with animate head nouns	…the guy that made that call
	Simple phrasal embedding in the noun phrase: nouns as premodifiers	…some really obscure cable channel
	Possessive nouns as premodifiers	Tobie's voice
	Of phrases as postmodifiers	**editor** of the food section
	Simple PPs as postmodifiers, especially with prepositions other than of when they have concrete/locative meanings	**house** in the suburbs
4	Nonfinite complement clauses controlled by adjectives	These will not be easy to obtain.
	Extraposed complement clauses	**It is clear** that much remains to be learned …
	Nonfinite relative clauses	…the **method** used here should suffice …
	More phrasal embedding in the NP = attributive adjectives, nouns as premodifiers	The prevalence of airway obstruction and self-reported disease status
	Simple PPs as postmodifiers, especially with prepositions other than *of* when they have abstract meanings	with half of the subjects in each age/instructional condition receiving each form
5	Preposition + nonfinite complement clause	The **idea of** using a Monte Carlo approach
	Complement clauses controlled by nouns	The **hypothesis** that female body weight was more variable
	Appositive noun phrases	The **CTBS** (the fourth edition of the test) was administered in 1997–1998.
	Extensive phrasal embedding in the NP: multiple prepositional phrases as postmodifiers, with levels of embedding	The [presence of layered [[structures] at the [[[borderline]] of cell territories]]]

Note. The bold and underlined parts mark the target grammatical structures.

Lexical Complexity

Lexical complexity has been interpreted in various terms of lexical richness, lexical density, lexical sophistication, lexical variation, lexical diversity, and so forth. These terms are not only oftentimes used interchangeably but also hierarchically and exclusively to each other (Yu, 2010). Such terminological confusion together with the fact that the same term can be conceptualized and operationalized differently from study to study reflects the multidimensionality of lexical complexity, which is perhaps even greater than that of syntactic complexity. According to Bulté and Housen (2012), lexical complexity has four dimensions of density, diversity, sophistication, and compositionality. This somewhat corresponds to Read's (2000) description of lexical richness.

While lexical density is typically defined as the proportion of content or lexical words to total or functional/grammatical words, lexical diversity has to do with the range of one's vocabulary and has traditionally been measured as type-token ratio (TTR) – the proportion of unique words to total words. However, TTR is sensitive to text length as it decreases in longer texts due to increasing word repetition (McKee et al., 2000; Richards, 1987). Alternatively, measures such as D (Malvern et al., 2004) and MTLD (McCarthy, 2005) have recently been proposed to mitigate the influence of text length, thus enabling the comparison among texts of different lengths. A higher D means a lexically more diverse text. The index is especially useful for gauging the lexical diversity of short texts under 1,000 words, but it is still affected by text length (Jarvis, 2002; McCarthy & Jarvis, 2007). In comparison, MTLD has been found to be less affected (Koizumi & In'nami, 2012) or not vary as a function of text length (McCarthy & Jarvis, 2010). The validity and reliability of these measures have been displayed in abundant research (e.g., Crossley et al., 2011; McCarthy & Jarvis, 2007, 2010; Yang & Kim, 2020; Yoon & Polio,

2017; Yu, 2010). McCarthy and Jarvis (2010) particularly suggest combining multiple measures (e.g., D and MTLD) because they appear to contribute unique lexical information.

Another lexical diversity estimate claimed to be able to control for text length effects is the Guiraud index (G), a mathematical transformation of the typical TTR. It is also claimed to be one of the most robust type-token measures (van Hout & Vermeer, 2007). With regard to lexical sophistication, it usually involves corpus-based frequency information of an L2 learner's (advanced) lexis. It is widely measured as average word length (Jarvis et al., 2003; Verspoor et al., 2012; Yoon & Polio, 2017), as longer words tend to be more sophisticated, and frequency-based type/token ratios like lexical frequency profile (Laufer & Nation, 1995). Besides the dimensions above, Bulté and Housen (2012) have added compositionality, which refers to the number of formal and semantic elements of lexical items, with operationalizations of morphemes/words and syllables/words ratios.

Several surveys have been conducted on the measurement of lexical complexity. Johnson's (2017) synthesis showed that task-based L2 writing researchers used from one to six different measures of lexical complexity in their studies (see Table 4). Johnson observed that only three of the included studies employed D and MTLD, which are newer, more sophisticated and less text-length reliant measures of lexical complexity. In similar reviews of studies examining the effects of cognitive task complexity on CALF (Bulté & Housen, 2012; Yang, 2014), 68% of the studies were found to cover lexical diversity measures. Only 36% and less than 10% of the studies took into account lexical density and lexical sophistication, respectively. Moreover, only one or two measures of lexical density and sophistication were employed.

Table 4

Lexical Complexity Measures in Task-Based L2 Writing Research (adapted from Johnson, 2017)

Category	Measure	Note
Lexical density	Lexical words/Total words	
	Lexical words/Function words	
Lexical diversity	MTLD	Measure of textual lexical diversity
	D	
	vocD	
	Pronouns/Noun phrase	
	Type-token ratio	
	Mean, segmental type-token ratio	
	Corrected type-token ratio	
	$Type^2/\sqrt{Token}$	
	Giraud's index	
Lexical sophistication	% of 1 K GSL words	The first 1,000 most frequent word families according to the general service list
	% of 2 K GSL words	The second 1,000 most frequent word families according to the general service list
	Words from beyond the 2 K list	
	% of AWL	Academic word list
	BNC 4 K words/100 words	The fourth 1,000 word families according to the British National Corpus (BNC; Davies, 2004)
	BNC 5 K words/100 words	The fifth 1,000 word families according to BNC
	CELEX rating	
	Log CELEX rating	
	Concreteness of Lexical words	

Automated text analysis tools such as Coh-Metrix, the Lexical Complexity Analyzer (LCA; Lu, 2012) and the Tool for the Automated Analysis of Lexical Diversity (TAALED; Kyle et al., 2021) as well as the Tool for the Automated Analysis of Lexical Sophistication (TAALES; Kyle & Crossley, 2015; Kyle et al., 2018) are also available for the measurement of lexical

complexity. Among the large array of linguistic indices computable by Coh-Metrix, many tap into lexical complexity, with D, MTLD, word length, and word frequency being notable measures (see Crossley et al., 2011; MacArthur et al., 2019; McNamara et al., 2010; Yoon, 2017). LCA and TAALED report 25 and 10 indices, respectively. The former embraces three aspects of lexical complexity (i.e., lexical density, diversity, and sophistication), whereas the latter's focus is on lexical diversity measures. Specialized in measuring lexical sophistication, TAALES provides 135 indices pertinent to word frequency, range, bigram and trigram frequency, academic language, and psycholinguistic word information. It was mainly developed to accommodate new and potentially important measures and to process a large number of texts in a reasonable amount of time. It is also validated with indices capable of explaining 47.5% of the variance in human judgements of lexical proficiency in L1 and L2 writing. The programs typically draw on wordlists from large, representative corpora such as COCA and BNC. It should be noted, however, that many measures incorporated in these programs are transformations of several core measures. In Bulté and Housen's (2012) survey of complexity measurement, an observation was made that most studies measuring L2 complexity only calculated a few measures. The mean number of measures used in the 40 surveyed studies is 2.7, and 22 of the studies used just one or two measures. Given the automated tools described here, it is now feasible to address this problem.

Using only a few measures to assess L2 complexity can be problematic for several reasons. Firstly, complexity is a multidimensional construct, and assessing it requires a comprehensive understanding of its various components. If only a few measures are used, important aspects of complexity may be overlooked. Secondly, the few measures that are often used to assess writing complexity may not adequately capture the range of difficulties that

learners may encounter when writing in an L2. Thirdly, relying on a limited set of measures may oversimplify the assessment of L2 complexity, leading to inaccurate or incomplete conclusions about learners' writing abilities. For instance, if only one or two measures are used, learners may receive misleading feedback on their writing, which may hinder their progress in the L2. Therefore, to provide a comprehensive and accurate assessment of L2 writing complexity, multiple measures should be used, covering various aspects of the writing process.

Accuracy

If CALF constructs are put on a scale, accuracy or correctness is likely "the most straightforward and internally consistent construct" (Housen & Kuiken 2009; Pallotti 2009, as cited in Housen et al., 2012, p. 4). Polio (2001) generally equates accuracy with the absence of errors, but it should be more precisely regarded as the extent to which L2 production deviates from a norm (Pallotti, 2009). Accuracy measurement or error identification thus poses the question of what counts as an error. In other words, whether errors are to be evaluated by NS norms or by English as a lingua franca norms must be made clear. Also, the extent to which errors hinder communication can help weigh errors, but it should be noted that, for example, a text with 10 errors not compromising communication is not more "accurate" than the same text with 10 errors hindering comprehension, but just more "understandable" or "communicatively effective" (Pallotti, 2009, p. 592). Another point of consideration is the type of error committed by learners. A text with 10 errors on subjunctives and conditionals is not more "accurate" than the same text with 10 errors on articles and pronouns, but just more "developed" or "advanced" (Pallotti, 2009, p. 592). Thus, researchers should be careful not to judge accuracy using measures related to other constructs such as intelligibility and development. How seriously researchers take these concerns into account while designing their studies has yet to be reported.

Just like Bulté and Housen's (2012) comment about the range of L2 complexity measures covered in individual studies, many task-based L2 writing studies include a very limited number of accuracy measures (Johnson, 2017). These studies tend to operationalize accuracy as some form of error-free units and/or error count per unit, with or without specifying error types (Johnson, 2017; Yang, 2014). Example measures are the percentage of error-free clauses/T-units and the number of errors per clause/T-unit. Holistic scales and qualitative analysis can also be used to measure accuracy, as reviewed by Polio (2001). Foster and Wigglesworth (2016) divide commonly used accuracy measures in L2 performance in two broad categories: local measures aiming to monitor the use of selected grammatical features and global measures focusing on determining the overall level of accuracy. After reviewing the measures, they came to the conclusion that global measures based on a syntactic unit can evaluate accuracy in L2 performance better than local measures, with the error-free clause being one of the strongest units of measurement.

Polio and Shea's (2014) study provides a more up-to-date and comprehensive review of accuracy measures in L2 writing research. Their results seem to resonate with other syntheses as they found holistic scales, error-free units, number of errors, number of specific error types, and error gravity were the types of accuracy measures included most frequently. What is most striking is that intra or interrater reliability was not reported for almost half of the 44 accuracy measures from the 35 studies they examined and that both types of reliability were reported for only four measures. Given the information gained from their literature review, Polio and Shea (2014) selected 10 measures and tested their reliability on a dataset from Michigan State University, which consists of 210 timed descriptive essays written in 30 minutes by Intensive English Program and English for Academic Purposes (EAP) students over the course of a

semester (each student wrote three essays). The chosen measures were holistic scores of language use, holistic scores of vocabulary, error-free T-units/total T-units, error-free clauses/total clauses, weighted error-free T-units/total T-units, number of errors per words, and number of verb phrase, preposition, article, and lexical errors per words. Both researchers coded all the data. Measure validity was also examined by checking correlations among the measures, change over time, and differences between the groups showing improvement and no improvement. All the measures achieved interrater reliability coefficients greater than .84 except for the specific error types: verb phrase errors (.65), preposition errors (.79), article errors (.80), and lexical errors (.54). The validity analyses suggest that there is not a measure of accuracy that is more valid than others. Since no measure really stands out in terms of reliability and validity, researchers should be informed of the pros and cons of each measure and combine different measures (Polio & Shea, 2014).

What is as important as measure selection is developing and reporting specific coding guidelines and reliability for replication and research rigidity. An example of this is Yoon and Polio's (2017) comparison of English as a Second Language (ESL) students' linguistic development between two written genres. The authors made conscious decisions to include syntactic, morphological, preposition, and spelling error types and exclude lexical errors as measures of accuracy. Importantly as well, the researchers provided readers with clear examples of each category and guidelines for coding these different error types in future studies. In their own study, they reached acceptable interrater reliability on all measures (syntactic errors = .84; morphological errors = .96; preposition errors = .92) with spelling errors counted using an online spell checker. It is apparent from their report that they were aware of the advantages of the included measures and had a clear picture of what constitutes an error and what does not.

Fluency

Fluency in L2 writing is the pace with which the L2 is written, but its meaning is usually interpreted as how nativelike the writing sounds (Polio, 2001). Wolfe-Quintero et al. (1998) defines L2 writing 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 structural units a writer is able to include in their writing within a particular period of time. (p. 25)

Reflecting on this definition, Polio (2001) particularly argued that fluency should be measured by counting the number of words produced in a given time because counting the number of other structural units such as clauses and T-units would penalize learners who write longer structures. Fluency thus seems easy to measure, especially with the assistance of word processing programs, but conceptually what constitutes a word is sometimes confusing. Moreover, the number of words may be related to various factors such as planning time and may not measure how quickly the writer writes (Abdel Latif, 2013). Polio (2001) also questioned the relation of fluency to writing quality and thus the role it plays in the writing process. This is probably part of the reason why writing fluency is not in the spotlight in CALF studies. However, it is important to measure fluency to understand how CALF measures interact with each other (Yoon & Polio, 2017).

In addition to measuring the amount of production based on some unit, fluency has been measured using holistic scales (Polio, 2001). In Abdel Latif's (2013) call for better understanding of writing fluency and how it should be measured, fluency measures were divided into two types: product-based measures relying on written texts despite how they were produced (e.g., changes made, composing rate, text quantity) and process-based measures drawing upon the online observation of writers' composing processes (e.g., pausing, length of rehearsed text, and length of translating episodes). Abdel Latif (2013) concluded that writing fluency could be validly measured by the length of writers' translating episodes or production units, which assesses real-time writing and is reflective of the cognitive characteristics of writing performance.

The Relationship Between CALF and L2 Writing Quality

Syntactic Complexity and L2 Writing Quality

Insights into what and how syntactic complexity features contribute to good writing are important for SLA theory, pedagogy, and assessment. Accordingly, one crucial question in the study of L2 writing complexity is about the relationship between syntactic complexity and writing quality, with the latter normally reflected in holistic or analytic essay scores given by human raters. A common hypothesis in this line of research is that learners of higher proficiency or with linguistic maturity have more control over complex syntactic structures and thus will write with more efficiency and flexibility (Ortega, 2015; Yang et al., 2015). This view holds that L2 writing quality is a function of language proficiency and assumes it to be positively correlated with syntactic complexity. However, researchers have revealed a more complex picture of the relationship due to different factors affecting the relationship and the variety of syntactic complexity measures used across studies.

In one of the first research syntheses of the relationship between syntactic complexity and L2 proficiency for college-level writing, Ortega (2003) focused on six syntactic complexity measures that were most frequently used in her sample of 21 cross-sectional studies: MLS, MLT, MLC, T/S, C/T, and DC/C. Her statistical analyses suggested the following critical magnitudes for between-proficiency level differences in syntactic complexity: 4.5 or more words per sentence (MLS), 2 or more words per T-unit (MLTU), slightly over 1 word per clause (MLC), and at least a 0.20 positive or negative difference in C/T. The amount of syntactic complexity was typically higher in ESL writing than EFL, and the ranges of observed complexity values were narrower for studies using holistic ratings than those relying on program levels to establish proficiency group differences. Yet, the extent to which instructional setting (ESL vs. EFL) and proficiency sample criterion (program levels vs. holistic ratings) affect the proposed magnitudes is unclear.

Later, Lu (2017) reviewed corpus-based L2 writing studies investigating the relationship between syntactic complexity and writing quality with automated tools for syntactic complexity analysis (see the article for a summary table of the syntactic complexity measures found to be predictive of or correlated with L2 writing quality). His conclusion was:

> Studies employing the Biber Tagger found that a number of clause- and phrase-level grammatical complexity features could distinguish high- and low-scored essays but that the co-occurrence patterns of grammatical complexity features are more predictive of writing quality than individual features. (p. 504)

Meanwhile, certain Coh-Metrix measures of syntactic pattern density have been found to be correlated with holistic ratings of writing quality. These measures are the average number of modifiers per noun phrase and normed rates of occurrence of infinitives, negations, verb phrases,

and prepositional phrases (Crossley & McNamara, 2014). A number of SCA measures have also been reported to strongly correlate with writing quality: MLS, MLC, MLTU, DC/C, and CN/C (Chen et al., 2014; Li, 2015). These studies particularly sampled Chinese EFL learners whose essays were rated using the College English Test holistic rubric. Missing in Lu's (2017) literature review are studies that adopt the more recent TAASSC. Kyle and Crossley (2018) used both the SCA (which computes traditional syntactic complexity indices) and TAASSC (which computes fine-grained clausal and phrasal indices) to analyze TOEFL independent essays. Their results showed that fine-grained indices of phrasal complexity (e.g., number of dependents per prepositional object) were more powerful than either traditional syntactic complexity indices (e.g., MLC) or fine-grained clausal complexity indices (e.g., number of subjects per clause) in predicting holistic essay scores. As Lu (2017) recapped, a positive relationship was found between syntactic complexity and L2 writing quality despite how automated complexity measures or holistic quality ratings were operationalized. It was also noted that more in-depth analysis of the reliability of human judgements and the distributions of complexity scores would be necessary to compare the magnitudes of correlations on particular measures.

A wide range of syntactic complexity measures has been reported to be strongly correlated with or even predictive of L2 writing quality, but more intricate mappings of the relationship between syntactic complexity and writing quality have seen it vary across genres and topics. One example is Beers and Nagy's (2009) study on 41 seventh and eighth graders' writing. In addition to examining the relationship of syntactic complexity with rated writing quality, the authors drew a comparison between two genres: narratives and argumentative essays, from which differences arose. In particular, the relationship between syntactic complexity and quality ratings varied from one genre to the other. Bivariate correlational analyses revealed a

positive correlation between words per clause and writing quality and a negative correlation between C/T and quality for argumentative essays. C/T, conversely, were positively correlated with writing quality for narratives. However, the writing was completed in the participants' L1 (i.e., English). In comparison, Qin and Uccelli (2016) examined 100 sixth to eleventh grade Chinese EFL learners' writing performance in the same genres, which was evaluated using two genre-specific holistic rating rubrics. Unlike Beers and Nagy's (2009) findings on L1 writing, Qin and Uccelli (2016) found words per clause to have positive and stronger correlations with writing quality in both argumentative essays and narratives. No significant correlation between C/T and quality was reported in either genre.

Another variable that may play a role in the relationship between syntactic complexity and writing quality is writing topic. Yang et al. (2015) discovered in the TOEFL independent writing of 190 graduate students that topic had no significant effect on MLS and MLT, but these global complexity features were positively and significantly correlated with essay scores across the two topics. At the local complexity levels, topic effects were significant and greater. More specifically, the topic requiring causal reasoning (importance of planning for the future) elicited a significantly higher amount of finite and non-finite subordination, whereas the other topic (whether personal appearance and fashion are overemphasized) involved significantly more elaboration at the finite clause level, through the use of more coordinate phrases and complex noun phrases. When it comes to writing quality, the local-level syntactic complexity features strongly correlated with essay scores for the future topic but did not for the appearance topic, although the future topic observed significantly fewer of these features. On the contrary, finite subordination strongly correlated with essay scores for the appearance topic but did not for the future topic, although the appearance topic observed significantly lower amount of finite

subordination. Non-finite subordination was used more in future essays and was also strongly correlated with essay scores. These findings suggest that the students who were able to use local-level complexity features in addition to topic-intrinsic features achieved higher scores, demonstrating their higher L2 writing ability and/or proficiency. Not only does Yang et al.'s (2015) work help paint a more complete picture of the impact various factors may have on the relationship between syntactic complexity and writing quality, but it also addresses previous studies' limitations by using a larger sample size and measuring syntactic complexity as a multi-dimensional construct.

Attempting to link syntactic growth in L2 writing to writing quality, Bulté and Housen (2014) analyzed 45 adult ESL learners' essays to determine indicators of writing development and quality. Of the 10 syntactic complexity measures selected, seven were progress-sensitive. The measures having the highest effect sizes were those based on average length of linguistic units: mean length of finite clause (d = 0.49), MLT (d = 0.47), and MLS (d = 0.44). Interestingly, these progress-sensitive measures did not match perfectly with the predictors of overall writing quality. For example, while complex sentence and subclause ratios correlated with subjective ratings of writing quality, compound sentence and coordinate clause ratios did not, although their scores increased significantly over time. Similarly, Crossley and McNamara (2014) concluded that the syntactic complexity measures showing progress may not overlap with those associated with higher essay scores. More specifically, they suggested that L2 learner growth is demonstrated by a stronger nominal style and phrasal features, whereas human judgements of L2 writing quality is better predicted by clausal complexity. In sum, Crossley and McNamara claimed that incidence of all clauses (i.e., Coh-Metrix's normalized incidence counts for matrix,

coordinated, and embedded clauses) is the only meaningful indicator of both L2 writing development and quality.

Clearly, much previous research has examined the ability of syntactic complexity measures to distinguish L2 proficiency levels and the role it plays in L2 writing development (e.g., Biber et al., 2011; Bulté & Housen, 2014; Lu, 2011; Ortega, 2003; Wolfe-Quintero et al., 1998; Yoon & Polio, 2017). Nevertheless, direct investigation of the relationship between syntactic complexity and writing quality with the latter recorded by holistic or analytic essay scores is still lacking (Yang et al., 2015), which can help better understand the role of syntax in contributing to overall writing quality. Meanwhile, Ortega (2015) assumes certain correlations of syntactic complexity measures with writing quality and asserts the usefulness of having a better understanding of which syntactic complexity measures contribute to good writing or supposedly high ratings of human judges from both developmental and educational perspectives.

Lexical Complexity and L2 Writing Quality

Like syntactic complexity, lexical complexity has been examined for its correlation with L2 writing quality. Measures of lexical complexity such as G and D have stood out as strong discriminators of different general proficiency and writing quality levels. Examining objective measures of 437 texts written by beginner and intermediate L2 Dutch learners of English from a dynamic usage-based perspective, Verspoor et al. (2012) concluded that G was the most useful in distinguishing between all adjacent levels compared to word length (lexical sophistication) and customized lexical frequency profile (lexical originality). In their study, the texts were holistically coded for six proficiency levels (0-5) and organized in the Common European Framework of Reference for Languages (CEFR) levels (A1.1-B1.2). Bulté and Housen (2014) had 90 ESL essays rated for overall writing quality based on content, organization, language use,

vocabulary, and mechanics, and again, found G to have the strongest correlation with subjective ratings of writing quality. Although the correlation between D and L2 writing quality was demonstrated as weak in Bulté and Housen's (2014) study, other research has shown opposite results.

In Crossley and McNamara's (2012b) comparison of the roles cohesion and linguistic sophistication play in predicting senior Hong Kong high school students' essay grades, lexical diversity, word familiarity, word frequency, and word meaningfulness accounted for most of the variance in the multiple regression model. The measure D alone (representing lexical diversity) accounted for 18% of the variance. Similar results have been generated in other studies such as Yang (2014) and Yu (2010). Yang (2014) found D to be strongly correlated with writing scores of EFL Chinese university students for narrative and expository-argumentative tasks, whereas Yu (2010) observed that compositions with a higher D had the tendency to receive a higher score across genders, L1s, writing purposes, and writing topics, with D accounting for roughly 11% of the variances in the overall quality ratings.

In addition to the G and D, several other lexical complexity measures are also able to discriminate L2 writing of different quality levels. In Yang's (2014) investigation, for example, lexical sophistication as determined by the proportion of sophisticated word types was strongly correlated with all four rhetorical tasks: narrative, expository, argumentative, and expository-argumentative. Other significant correlations with proficiency or writing quality include MTLD (lexical diversity; McNamara et al., 2010), the ratio of lexical words to the total number of words (lexical density; Kim, 2014), the ratio of sophisticated verbs to the total number of verbs (lexical sophistication; Kim, 2014), and the number of different words as well as the ratio of different lexical words to the total number of lexical words (lexical variation; Engber, 1995; Kim, 2014).

Moreover, Johnson et al. (2012) selected five measures of lexical complexity that correlate with holistic ratings of L2 writing in their examination of the impact pre-task planning has on L2 writing fluency, grammatical complexity, and lexical complexity. These measures are: (1) MTLD, (2) the ratio of pronouns to noun phrases, (3) the incidence of personal pronouns normed to 1,000 words, (4) the mean frequency rating with which the content words in a text appear in the English language according to the COBUILD English language corpus, and (5) the normed frequency (per 100 words) of word types from the fourth and fifth most frequent word families according to the BNC (Coniam, 1999; Engber, 1995; Grant & Ginther, 2000; Jarvis et al., 2003; Lemmouh, 2008). Overall, significant correlations between lexical complexity measures and L2 writing quality have been observed in previous research, but it seems that researchers have tended to overlook the potential influences of independent variables such as L1 and writing genres and topics on the relationships, although lexical complexity has been shown to vary according to these factors (Yang, 2014; Yoon, 2017; Yoon & Polio, 2017; Yu, 2010).

Accuracy and L2 Writing Quality

Accuracy measures have been included in L2 writing research to investigate the effects of written corrective feedback, the effects of planning, the effects of task complexity, the difference between individual and collaborative writing, and change over time (Polio & Shea, 2014). Direct investigations into their relationship with L2 writing quality is undoubtedly scarce, although abundant research on L2 writing errors has been carried out, especially in EAP contexts, to identify learners' needs and help them achieve better accuracy (e.g., Chuang & Nesi, 2006; Romano, 2019; Singh et al., 2017; Wee et al., 2010). Theoretically, it is often assumed that higher linguistic accuracy results in better writing quality as reflected by writing scores.

Wolfe-Quintero et al. (1998) examined studies measuring accuracy among different proficiency groups and targeting the correlations between accuracy measures and holistic measures of writing proficiency or essay quality. In some of the studies they examined, accuracy measures were correlated with proficiency measures, but the studies did not measure proficiency or essay quality consistently. In others, accuracy measures were not correlated with proficiency level but holistic measures of writing proficiency or quality. In Verspoor et al.'s (2012) study, the relative number of errors per text or accuracy rate was not consistent in distinguishing holistically coded proficiency levels. More specifically, lexical errors discriminated between the two lowest levels, and verb use errors discriminated between the third and fourth levels. Thus, they found more specific measures to have more discriminatory power than the more general measure of number of errors. Their results, together with Polio and Shea's (2014) discussion, raise the question of whether there exists a universal measure(s) of accuracy that can be applied across proficiency levels and contexts.

Fluency and L2 Writing Quality

Not much has been written about the relationship between fluency and L2 writing quality. Quality may suffer with increased writing speed, causing a negative correlation between fluency and quality (Polio, 2001). However, several studies have shown that the more fluently L2 learners write the higher quality their essays are, and that fluency can predict quality (Friginal et al., 2014; Jarvis et al., 2003). Instead of assuming a linear relationship between CALF measures and quality ratings, Jarvis et al. (2003) took a different approach and performed cluster analyses to explore multiple profiles of highly rated timed essays in terms of CALF. What the researchers mean by *multiple profiles* is that one profile of highly rated essays may be longer than average, has lower-than-average mean word length, and has lower-than-average lexical diversity, whereas

another profile may be average in length, has above-average mean word length, and has above-average lexical diversity. Their cluster analyses revealed W/Tx as a dominantly strong grouping factor: all profiles or clusters of highly rated texts displayed longer-than-average text length with short-but-higher-rated texts and long-but-lower-rated texts being extremely rare. Following this approach, Friginal et al. (2014) found highly rated compositions could be clustered into six different linguistic profiles across NS and NNS groups. The profiles overlapped when four of them had high mean Z scores for W/Tx. Writing quality was assessed by means of the internet-based TOEFL (TOEFL iBT) rubric on a scale of 0-5 in Friginal et al. (2014) and holistic scales of 1-10 and 1-6 in Jarvis et al. (2003), but interrater reliability of essay scoring was reported in neither of the studies. Clearly, given the scarcity of research involving the relationship between fluency and writing quality, this is an area in need of further investigation.

The Relationship Between CALF and L2 Writers' L1

A common comparison in research on academic writing is between NS and NNS writers (e.g., Ädel & Erman, 2012; Cao & Xiao, 2013; Mansourizadeh & Ahmad, 2011; Salazar, 2014). Applying the same comparison to written syntactic complexity, Ai and Lu (2013) analyzed 400 essays written by English major students from the Written English Corpus of Chinese Learners Version 2.0 and 200 essays written by American university students from LOCNESS. The NNS writers were divided into two groups of high and low proficiency levels, and a set of 10 syntactic complexity measures was employed. The analysis resulted in significant differences in all syntactic complexity dimensions (i.e., length of production unit, amount of subordination, amount of coordination, and degree of phrasal sophistication) between NS and NNS students. The same patterns of difference between NS and NNS writing were observed for both higher and

lower NNS proficiency groups, except that the more proficient NNS students were significantly closer to the NS students in terms of production unit length and degree of phrasal sophistication.

Ai and Lu's (2013) study undoubtedly helps validate the use of syntactic complexity measures to gauge differences in language proficiency and warns language instructors and program directors of the gap in syntactic complexity between NS and NNS groups of students so that they can devise suitable treatments for the latter. However, given the learner sample used in their study, these implications might be applicable to Chinese students only. Little is known about inter-NNS or L1 differences in syntactic complexity. Advancing this line of research is urgent to cope with the diversity in the English-speaking population.

Previous research has provided strong evidence that writers' L1 background can influence the study of syntactic complexity or syntactic complexity itself in L2 writing (Crossley & McNamara, 2012a; Jarvis & Crossley, 2012; Lu & Ai, 2015). An example of such evidence is Lu and Ai's (2015) exploration of L1-related differences in syntactic complexity in L2 writing of college students from the following L1 groups: Bulgarian, Chinese, French, German, Japanese, Russian, and Tswana. Lu and Ai (2015) collected 1,400 argumentative essays written by college-level EFL learners of seven L1 backgrounds from ICLE 2.0 and 200 essays produced by native English speaking university students in the U.S. from LOCNESS. The essays were analyzed using SCA. While independent samples *t*-tests indicated that the NS and NNS groups differed significantly in MLC, CN/C, and CN/T, one-way ANOVAs suggested significant differences between the NS group and at least one NNS group in all the 14 measures of syntactic complexity examined. Significant differences were also found between the NNS groups when they were compared to the NS group. For example, the French, German, and Russian groups all

demonstrated significantly more sentential coordination than the NS group, which was not applicable to any of the Chinese, Japanese, and Tswana groups.

Nevertheless, Lu and Ai (2015) pointed out that the few studies comparing syntactic complexity in NS and NNS writing "did not treat learners' L1 background as an independent variable but either looked at a homogeneous L1 group or treated all NNS learners as one group" (p. 17). The same case is true for lexical complexity. As an example, Crossley and McNamara (2009) compared argumentative essays written by Spanish learners of English and English NSs using 10 lexical variables from Coh-Metrix (e.g., word frequency, meaningfulness, hypernymy, polysemy). They came to the conclusion that the L1 writers were lexically more proficient than their L2 peers, supporting previous findings that L2 writers display less lexical variation and sophistication (Linnarud, 1986; Nakamaru, 2010).

More recently, Eckstein and Ferris (2018) used LCA to investigate lexical differences between English NSs and NNSs who were participating in the first-year composition program with the NNSs coming from a variety of L1 backgrounds. Differences in lexical variety, especially verb variation, were found despite similarities in lexical density and lexical sophistication. These findings, unfortunately, are not L1-specific, and studies attempting to distinguish L2 writers of different L1s based on linguistic complexity are still scarce (e.g., Crossley & McNamara, 2012a; Jarvis, 2002). Accordingly, it is integral to examine how another factor, L1 background, functions in the complexity - writing quality relationship. It is also important to compare syntactic complexity and lexical complexity measures to see what types of measures might be more powerful in predicting writing quality and representing certain L1 backgrounds.

Similar to the assumption about the relationship between accuracy and L2 writing quality, it can be assumed that native writers are more linguistically accurate than nonnative writers. Barrot and Gabinete (2019) investigated CALF differences in ESL and EFL learners' argumentative writing. The results indicated, for example, that Filipino and Singaporean ESL learners wrote more accurate essays than the EFL group with a medium effect size. Interestingly, the Pakistani and Hong Kong participants, who were also ESL learners, did not demonstrate the same or even a similar pattern. The findings, however, do not tell much about the specific areas where the L1 groups differ from each other because converted, generic measures of accuracy such as error-free clauses of all clauses and the proportion of error-free T-units of all T-units were adopted in the study without additional information on the coding scheme or error frequency counts, even though the authors stated the learners' L1 backgrounds might have played a crucial role in CALF differences in L2 writing.

Eckstein and Ferris (2018) previously made a similar comparison between L1 and L2 students participating in a 10-week first-year composition course in the U.S. They coded the students' errors based on nine major categories: punctuation, mechanics, nouns/noun phrases, subject-verb agreement, verbs/verb phrases, sentence structure, word form, pronoun usage, and incorrect word choice. Different from Barrot and Gabinete (2019), the researchers reported frequency and type of language errors, showing that there were more errors in the L2 texts in every category. Unfortunately, in this study, the L2 students coming from different L1 backgrounds were treated as one holistic NNS group. Another issue in previous research on L2 writing accuracy is what counts as an error varies from study to study, and many studies reported very few accuracy measures (Johnson, 2017). Thus, a wide range of accuracy measures need to be covered, facilitating direct comparisons between the measures.

The Current Study

In summary, previous findings on L2 writing complexity and accuracy show that if L1 is ignored in research design, errors may occur in measuring linguistic features, distorting the significance of any results found, and knowledge of how L1 may affect syntactic (and lexical) complexity in L2 writing is lacking (Lu & Ai, 2015). Even less is known about the interactions between L1 background, CALF, and L2 writing quality altogether. The current study, therefore, aims to fill such research gaps by conceptualizing CALF as multidimensional constructs of language proficiency and examining them in terms of L1 and writing quality. It particularly seeks to investigate the following questions:

1. To what extent do CALF measures vary across L1 backgrounds in each score level?
2. To what extent do CALF measures vary across score levels in each L1 group?
3. To what extent can CALF measures and L1 backgrounds predict score levels?

With the research questions above, the study hopes to contribute to L2 writing research and pedagogy in several ways. It will first add to the growing body of research on L1 differences in L2 writing and help researchers determine whether L1 should be controlled for in different stages of their research. It will also shed light on previous findings that have reported on the relationship between complexity, accuracy, and writing quality regardless of potential L1 influences. In terms of L2 pedagogy, the study has the potential to inform language teachers about which linguistic features they should direct their students' attention to in order to improve performance, depending on the students' L1.

CHAPTER THREE: METHODOLOGY

The Corpus

This dissertation drew on TOEFL11 – a corpus of 12,100 TOEFL iBT essays written by test takers in 2006-2007. TOEFL is a widely accepted standardized test that measures the four academic English skills (reading, listening, speaking, and writing) of NNSs who wish to enroll in English-speaking universities, especially in the U.S. It is delivered by computer in secure test centers around the world and takes about four hours to complete (all four sections). The writing section of the test has a time limit of 50 minutes. It requires test takers to perform two writing tasks: an integrated writing task (20 minutes) and an independent writing task (30 minutes). While the former asks students to read a short passage and listen to a short lecture before writing in response to what they read and listened to, the latter asks them to write an essay based on personal experience or opinion in response to a writing topic. Only essays produced from the independent task contributed to the TOEFL11 corpus.

TOEFL11 is essentially a corpus of high-stakes essay writing. The driving force behind its birth was the first native language identification (NLI) shared task (Tetreault et al., 2013), in which many research teams competed to build statistical models to differentiate essays written in different languages. Nevertheless, the corpus is expected to be used in automated essay scoring, automated grammatical error detection and correction, corpus linguistic analyses of linguistic features across L1s, and cross-genre comparisons of writing as well (Blanchard et al., 2013). It is thus interesting to see its data analyzed under the current study's goals of linking CALF, writing quality, and writers' L1. Moreover, when compared to other corpora like ICLE, TOEFL11 has

certain advantages of not only containing a large set of essays but also having an even distribution of essay topics and consistent character encodings as well as annotations across various L1s (Blanchard et al., 2013), making it appropriate for the current study.

The sampling of TOEFL essays for TOEFL11 involves eight prompts and 11 L1s: Arabic, Chinese, French, German, Hindi, Italian, Japanese, Korean, Spanish, Telugu, and Turkish. Table 5 shows the number of essays from each L1 for each topic in the corpus. The L1s belong to seven language families: Romance (French, Italian, Spanish), Germanic (German), Indo-Iranian (Hindi), Altaic (Japanese, Korean, Turkish), Sino-Tibetan (Chinese), Afro-Asiatic (Arabic), and Dravidian (Telugu), among which Romance, Germanic, and Indo-Iranian are all Indo-European (see Figure 2 for the taxonomy of language families in TOEFL11). Given the possible presence of topic effects (He & Shi, 2012; Hinkel, 2009; Tedick, 1990; Yang et al, 2015; Yoon, 2017), only the essays responding to Prompt 8, which has a high number of essays and a fairly even distribution of essays across L1s, were included for analysis. The prompt asks test takers whether they agree with the following statement and to use reasons and examples to support their answer: "Successful people try new things and take risks rather than only doing what they already know how to do well".

Table 5

Number of Essays Per Language Per Prompt (adapted from Blanchard et al., 2013)

Language	Prompt 1	Prompt 2	Prompt 3	Prompt 4	Prompt 5	Prompt 6	Prompt 7	Prompt 8	Total
Arabic	138	137	138	139	136	133	138	141	1,100
Chinese	140	141	126	140	134	141	139	139	1,100
French	158	160	87	156	160	68	151	160	1,100
German	155	154	157	151	150	28	152	153	1,100
Hindi	161	162	163	86	156	53	158	161	1,100
Italian	173	89	138	187	187	12	173	141	1,100
Japanese	116	142	140	138	138	142	141	143	1,100
Korean	140	133	136	128	137	142	141	143	1,100
Spanish	141	133	54	159	134	157	160	162	1,100
Telugu	165	166	167	55	169	41	166	171	1,100
Turkish	169	145	90	170	147	43	167	169	1,100
Total	1,656	1,562	1,396	1,509	1,648	960	1,686	1,683	12,100

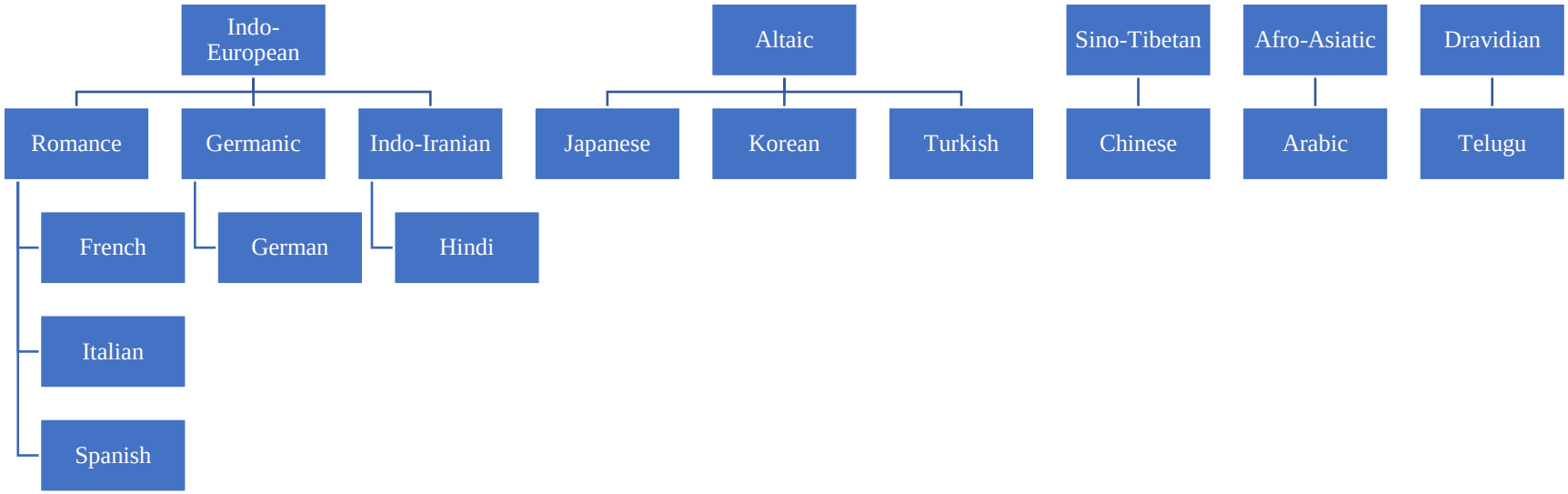

Figure 2. TOEFL11 Language Families (adapted from Blanchard et al., 2013).

TOEFL essays are scored on a 5-point scale with distinct rubrics for the independent and integrated writing tasks (https://www.ets.org/s/toefl/pdf/toefl_writing_rubrics.pdf) by both artificial intelligence (AI) scoring and certified human raters. However, all TOEFL11 essays were scored independently by two human raters because AI scoring was not put into use by the ETS until 2008. Although the scoring intra and interrater reliability were not reported for TOEFL11 specifically, TOEFL scores are deemed reliable given the strict procedures and guidelines used to achieve score reliability and comparability (ETS, 2020). In the corpus, the original 5-point scale scores were collapsed into a 3-point scale: low (essays scoring between 1.0 and 2.0), medium (2.5-3.5), and high (4.0-5.0). The current study operationalized writing quality by this scale. Table 6 shows the number of essays from each L1 for each score level. The length of the included essays ranges from 2 to 591 words, with an average of 314 word tokens per essay (*SD* = 72).

Table 6

Distribution of Essays Across L1s and Score Levels

L1	Low	Medium	High	Total
Arabic	41	72	28	141
Chinese	20	102	17	139
French	12	89	59	160
German	3	57	93	153
Hindi	8	79	74	161
Italian	14	85	42	141
Japanese	22	93	28	143
Korean	32	77	34	143
Spanish	14	86	62	162
Telugu	20	109	42	171
Turkish	12	103	54	169
Total	198	952	533	1,683

CALF Measures

Syntactic Complexity Measures

As suggested in the literature review, a variety of indices were automatically calculated for syntactic complexity to cover the four areas of overall complexity, complexity by subordination, complexity by subclausal or phrasal elaboration, and complexity by coordination (see Table 7). SCA (https://aihaiyang.com/software/SCA/) meets this need for multidimensionality with its ability to generate 14 indices of syntactic complexity, but it can only analyze up to 30 text files at a time using the batch mode of the web version. Thus, TAASSC, which also reports on the 14 SCA indices and can process texts in folders, was used to compute the indices.

Table 7

Syntactic Complexity Measures

Category	Measure	Label	Tool
Sentential complexity	Mean length of sentence	MLS	TAASSC
	Number of T-units per sentence	T/S	TAASSC
	Number of clauses per sentence	C/S	TAASSC
T-unit complexity	Mean length of T-unit	MLT	TAASSC
	Number of clauses per T-unit	C/T	TAASSC
	Complex T-unit ratio	CT/T	TAASSC
	Number of dependent clauses per T-unit	DC/T	TAASSC
	Number of coordinate phrases per T-unit	CP/T	TAASSC
	Number of complex nominals per T-unit	CN/T	TAASSC
	Number of verb phrases per T-unit	VP/T	TAASSC
Clausal complexity	Mean length of clause	MLC	TAASSC
	Number of dependent clauses per clause	DC/C	TAASSC
	Number of coordinate phrases per clause	CP/C	TAASSC
	Number of complex nominals per clause	CN/C	TAASSC

Lexical Complexity Measures

Due to its multidimensional nature, lexical complexity was captured through the sub-constructs of lexical density, lexical diversity, and lexical sophistication. Lexical density was measured by means of LD, lexical diversity by D and MTLD, and lexical sophistication by means of LS1 and LS2. I used TAALED to compute LD and MTLD, LCA to compute LS1 and LS2, and the Computerized Language Analysis (CLAN; MacWhinney, 2000) to compute D. CLAN is a software program developed for the analysis of natural language data, specifically language produced by children and L2 learners. It provides a range of tools for text, audio, and statistical analyses and is deemed a reliable tool for analyzing language data as it has been extensively used and tested in linguistic research. Table 8 lists the five specific lexical complexity measures and their respective computing systems.

Table 8

Lexical Complexity Measures

Category	Measure	Label	Tool
Lexical density	Lexical density	LD	TAALED
Lexical diversity	Index of lexical diversity	D	CLAN
	Measure of textual lexical diversity	MTLD	TAALED
Lexical sophistication	Lexical sophistication-I	LS1	LCA
	Lexical sophistication-II	LS2	LCA

Accuracy Measures

Linguistic accuracy was reported as the number of syntactic, morphological, preposition, spelling, and total errors per 100 words, following Yoon and Polio's (2017) guidelines. Syntactic errors include incorrect word order, sentence fragments, run-on sentences and comma splices, missing constituents, extra verbs or subjects in a clause, and infelicitous uses of relative clauses. Morphological errors include incorrect uses of word form, subject-verb agreement, plurals, genitives, articles, double negatives, wrong pronouns in terms of gender or case, and verb form problems. Finally, preposition errors include missing, extra, or wrong prepositions. Additionally, a new error type called "incomprehensible" was devised from the error coding process. Incomprehensible errors are instances where the intended meaning is obscured by a single error or a series of errors, which in turn encumbers reliable error identification and classification. This type of error was thus added to preserve potentially valuable information and prevent coders from guessing. Detailed examples of the error types are provided in Appendix A. Lexical errors were not coded because of its low interrater reliability (Polio & Shea, 2014). It was also beyond the bounds of possibility to mark every type of error available due to time and energy constraints. For instance, errors on the discourse level such as periphrastic-topic constructions and use of *it* as discourse deixis (Chan, 2010) had to be excluded. These errors also seem to be more subtle and are thus likely to affect reliability.

Because of the sheer volume of essays, only a sample was hand-coded for linguistic accuracy. This sample is comprised of approximately 30 essays from each L1 group, totaling 329 essays or approximately 20% of the entire dataset. It is distributed as evenly as possible across the three score levels. Specifically, for every 30 essays selected from each L1, 10 were randomly selected from the low level, 10 from the medium level, and 10 from the high level. Since L1

German speakers produced only three low essays and L1 Hindi speakers produced only eight, more essays were selected from the medium and high levels for these two groups. Table 9 shows the distribution of the essays hand-coded for accuracy across L1s and score levels.

Table 9

Distribution of Accuracy Sample

L1	Low	Medium	High	Total
Arabic	10	10	10	30
Chinese	10	10	10	30
French	10	10	10	30
German	3	13	13	29
Hindi	8	11	11	30
Italian	10	10	10	30
Japanese	10	10	10	30
Korean	10	10	10	30
Spanish	10	10	10	30
Telugu	10	10	10	30
Turkish	10	10	10	30
Total	101	114	114	329

Together with another rater, I coded over 10% of the accuracy sample (i.e., 35 essays) for interrater reliability, with 10% of the sample being a common benchmark that is frequently used across various domains of L1 and L2 writing research (e.g., Casal & Kessler, 2020; Johnson, 2017). The second rater is a Ph.D. student in Linguistics and Applied Language Studies with

extensive experience in teaching L2 writing. They were trained with a set of essays that was not part of the accuracy sample until feeling comfortable with the coding task. With acceptable reliability obtained for each accuracy measure (incomprehensible = .72, syntactic = .86, morphological = .95, preposition = .91, spelling = .98), I proceeded to code the rest of the sample (n = 294) alone. Unlike previous studies (e.g., Yoon & Polio, 2017), spelling errors were counted manually in the current study because spell checkers like Microsoft Word's checks for misspellings too rigidly.

Fluency Measure

Since composing time was held constant for all writers (i.e., 30 minutes), fluency was simply operationalized as text length or the total number of words written in a text (Amiryousefi, 2016; Barrot & Agdeppa, 2021; Polio, 2001; Wigglesworth & Storch, 2009; Yang, 2014). The number of clauses, T-units, or sentences was not used because it is more likely to reflect other aspects of writing rather than fluency (Abdel Latif, 2013; Polio, 2001). The fact that the essays were timed minimizes the possible effect of planning time on writing fluency. However, information regarding other factors related to topic familiarity and composing processes is not available due to the nature of the data. Thus, although the length of translating episodes or production units written between pauses is recommended as a valid process-based measure of writing fluency (Abdel Latif, 2013), it could not be used in the current study. To count the number of words in each essay, TAASSC was used. It is worth mentioning that the tool considers contractions such as *aren't*, *I'm*, and *that's* as two words.

Data Analysis

First, 1,683 TOEFL essays (in the form of text files) were analyzed by the automated tools of CLAN, LCA, TAALED, and TAASSC for complexity and fluency indices. However,

only 1,675 files were analyzed by LCA for two lexical sophistication indices because the software only takes texts that have a minimum of 50 words as input. Eight files did not meet this requirement and were thus excluded from the lexical sophistication analysis. The tools output multiple comma-separated values (CSV) files. Manual counts of errors for 329 essays were also entered into a spreadsheet. These CSV files and spreadsheet were subsequently imported into the R environment (R Core Team, 2022) for further analysis.

Research Question 1 (RQ1) regarding the effects of L1 on CALF in each score level and Research Question 2 (RQ2) regarding the effects of score level or writing quality on CALF in each L1 group were addressed by performing one-way ANOVAs. One-way ANOVA tests are used to determine whether there are any statistically significant differences between the means of three or more independent, unrelated groups (which in this case are 11 L1 groups and three score levels). For RQ1, one test was run for each CALF measure in each score level with the measure as the dependent variable and L1 as the independent variable. For RQ2, one test was run for each measure in each L1 group with score level as the independent variable.

Prior to the main analyses, the assumptions of normality and homogeneity of variances were checked. The normality assumption was checked using Shapiro-Wilk tests (α = .05) and double-checked using skewness and kurtosis of the distribution. More specifically, absolute z-scores of skewness and kurtosis were calculated to determine whether the distribution of each subset is non-normal (e.g., z-scores of either skewness or kurtosis over 3.29 for medium-sized samples [50 < n < 300]; Kim, 2013). Homogeneity of variance was checked using Levene's tests (α = .05). When a subset was not normally distributed, the non-parametric Kruskal-Wallis H test was used. In case of unequal variances, Welch ANOVAs were carried out. For any statistically significant effects that were found, the analysis was followed up with appropriate post hoc tests

(i.e., Tukey-Kramer tests for ANOVAs, and Games-Howell for Kruskal-Wallis and Welch ANOVAs). Games-Howell tests were used as post hoc tests for both Kruskal-Wallis tests and Welch ANOVAs because they are not only common but also powerful (Midway et al., 2020; Sauder & DeMars, 2019). Recommended effect sizes were also calculated for the main effects: omega-squared (ω^2) and adjusted omega-squared (adj. ω^2) for ANOVA and Welch ANOVA, respectively, and epsilon-squared (ε^2) for Kruskal-Wallis (Tomczak & Tomczak, 2014; Yigit & Mendes, 2018).

For Research Question 3 (RQ3), multinomial logistic regression (or simply multinomial regression) was performed to test the predictive power of CALF indices and L1 on score level, with score level being the dependent, categorical variable (three levels) and CALF indices and L1 (11 levels) being the independent variables or predictors. Two separate regression models were fitted: one with syntactic complexity, lexical complexity, and fluency measures using the full dataset (1,683 essays) and one with accuracy measures using the hand-coded subset (329 essays). The models were fitted by taking the "one vs. rest" approach, where the odds of each outcome are modelled against all other outcomes, using the R package *polytomous* (Arppe, 2013; Levshina, 2015). McFadden's pseudo R^2 was used to evaluate the goodness-of-fit of these models, with values ranging from 0.2 to 0.4 (corresponding to 0.7 and 0.9 in linear regression; Louviere et al., 2000) indicating a very good fit (Levshina, 2015).

Before the models were fitted, one-way ANOVA results and Pearson's correlation coefficients were examined to determine potential predictors and avoid multicollinearity. If certain CALF measures showed significant differences between score levels regardless of L1 and were not highly correlated with each other ($r < .70$; Mostafa & Crossley, 2020), they were included in the models. All analyses were performed using R (R Core Team, 2022). The R code

used to produce the results is available via the Open Science Framework (https://osf.io/e24vn/?view_only=564293bf71174df994b68f09c20de581).

CHAPTER FOUR: RESULTS

This study examined the impact L1 has on CALF measures of L2 writing and the abilities of CALF to differentiate and predict score levels. A total of 1,683 EFL essays were analyzed. The essays were written by 11 different L1 groups and were categorized into three levels of writing quality: low, medium, and high (see Table 6 for a summary of the dataset).

RQ1. To What Extent do CALF Measures Vary Across L1 Backgrounds in Each Score Level?

Table 10 shows that in each score level, CALF measures varied significantly across L1 backgrounds (17/26 measures for low, 23/26 for medium, and 22/26 for high levels). In the medium and high levels, in particular, L1 had a significant effect on all the measures of syntactic complexity and lexical complexity.

Low Level

In the low level of writing scores, all syntactic complexity measures but CT/T, CP/C, and CN/C varied significantly across L1s with small to medium effect sizes. Pairwise comparisons revealed that MLS, C/S, MLT, CN/T, MLC, and DC/C differentiated from one to three L1 pairs. Many of these pairs contained Korean. Korean and Telugu differed from each other in five measures (see Table 11).

As for lexical complexity, LD differentiated Arabic from Japanese ($p < .001$), Korean ($p < .001$), and Telugu ($p < .001$) and Korean from French ($p = .01$) and Italian ($p = .03$). Although MTLD and LS2 also varied significantly across L1s, they failed to differentiate any L1 pairs. In terms of accuracy, morphological and total errors differed significantly across L1s in this score

level, but only the number of morphological errors differentiated L1s, namely between Chinese and German (p = .04) and between German and Telugu (p = .048).

In the low score level, W/Tx discriminated six L1 pairs: Chinese-Korean (p = .03), French-Hindi (p = .01), French-Telugu (p = .01), Hindi-Korean (p = .01), Hindi-Turkish (p = .04), and Korean-Telugu (p < .001).

Table 10

Between-L1 Differences in CALF in Each Score Level

Measures	Low		Medium		High	
	p	ES	*p*	ES	*p*	ES
Syntactic complexity						
MLS	**< .001**	0.23	**< .001**	0.20	**< .001**	0.14
T/S	**.009**	0.09	**< .001**	0.08	**.003**	0.03
C/S	**< .001**	0.17	**< .001**	0.19	**< .001**	0.16
MLT	**< .001**	0.17	**< .001**	0.17	**< .001**	0.09
C/T	**.004**	0.10	**< .001**	0.15	**< .001**	0.11
CT/T	.25	0.01	**< .001**	0.07	**< .001**	0.05
DC/T	**.006**	0.10	**< .001**	0.14	**< .001**	0.11
CP/T	**.02**	0.08	**< .001**	0.10	**< .001**	0.06
CN/T	**.002**	0.12	**< .001**	0.17	**< .001**	0.08
VP/T	**.003**	0.11	**< .001**	0.15	**< .001**	0.11
MLC	**< .001**	0.16	**< .001**	0.09	**< .001**	0.06
DC/C	**.03**	0.06	**< .001**	0.08	**< .001**	0.09
CP/C	.60	0.04	**< .001**	0.05	**.007**	0.05
CN/C	.10	0.08	**< .001**	0.09	**< .001**	0.12
Lexical complexity						
LD	**< .001**	0.19	**< .001**	0.09	**< .001**	0.09
D	.12	0.03	**< .001**	0.04	**.03**	0.02
MTLD	**.048**	0.04	**< .001**	0.05	**.003**	0.05
LS1	.10	0.08	**< .001**	0.10	**< .001**	0.12
LS2	**.006**	0.12	**< .001**	0.14	**< .001**	0.16
Accuracy						
Inc	.30	0.12	.40	0.09	.80	0.06
Morph	**.01**	0.22	**.02**	0.19	**< .001**	0.23
Prep	.20	0.14	**.01**	0.14	**.005**	0.22
Spell	.05	0.18	.20	0.12	.50	0.08
Synt	.06	0.18	.20	0.12	.05	0.16
Total	**.007**	0.14	**.049**	0.08	**< .001**	0.27
Fluency						
W/Tx	**< .001**	0.14	**.004**	0.03	.05	0.03

Note. Significant level $p < 0.05$. Significant p values are in bold. MLS = mean length of sentence, T/S = number of T-units per sentence, C/S = number of clauses per sentence, MLT = mean length of T-unit, C/T = number of clauses per T-unit, CT/T = complex T-unit ratio, DC/T = number of dependent clauses per T-unit, CP/T = number of coordinate phrases per T-unit, CN/T = number of complex nominals per T-unit, VP/T = number of verb phrases per T-unit, MLC = mean length of clause, DC/C = number of dependent clauses per clause, CP/C = number of coordinate phrases per clause, CN/C = number of complex nominals per clause. Inc = incomprehensible errors, morph = morphological errors, prep = preposition errors, spell = spelling errors, synt = syntactic errors, total = total number of errors. LD = lexical density, D = index of lexical diversity, MTLD = measure of textual lexical diversity, LS1 = lexical sophistication-I, LS2 = lexical sophistication-II. W/Tx = number of words per text.

Table 11

Between-L1 Differences in Syntactic Complexity in Low Score Level

	Arabic	Chinese	French	German	Hindi	Italian	Japanese	Korean	Spanish	Telugu	Turkish
Arabic	-							MLT (p = .03)			
Chinese		-									
French			-								
German				-					MLC (p = .005)	CN/T (p = .044) MLC (p = .02)	
Hindi					-						
Italian						-		MLS (p = .03)			
Japanese							-				
Korean								-	MLT (p = .007)	MLS (p = .02) C/S (p = .01) MLT (p = .03) CN/T (p = .03) DC/C (p = .03)	
Spanish									-		
Telugu										-	
Turkish											-

Medium Level

As mentioned above, in the medium and high levels, all syntactic complexity and lexical complexity measures varied significantly across L1s. Due to their large numbers, the specific L1s that differed significantly from each other in syntactic complexity in these two levels are encapsulated in Appendix B. In the medium level, each syntactic complexity measure distinguished at least nine pairs of L1s. For example, VP/T differed significantly for 28 out of 55 possible pairs. VP/T was also the syntactic complexity measure that distinguished the most L1 pairs in this level. Syntactic complexity measures seem prominent in differentiating Telugu from other L1s as this language background appeared in 93 comparison pairs across the measures, followed by Korean (77 pairs).

Table 12 presents the differences in lexical complexity between specific L1s in the medium score level. As it can be seen in the table, LS2 separated the most L1 pairs (25). Arabic appeared in the most comparison pairs (24). It differed from every other L1 in at least one lexical complexity measure. Accuracy measures such as morphological, preposition, and total errors also varied significantly across L1s in the medium level. Only the number of morphological errors distinguished specific L1s: German-Telugu ($p = .03$). As for fluency, W/Tx distinguished between Arabic-Hindi ($p = .002$), French-Hindi ($p = .02$), Hindi-Italian ($p < .001$), Hindi-Japanese ($p = .01$), Hindi-Korean ($p = .02$), and Hindi-Spanish ($p = .04$) in this level.

Table 12

Between-L1 Differences in Lexical Complexity in Medium Score Level

	Arabic	Chinese	French	German	Hindi	Italian	Japanese	Korean	Spanish	Telugu	Turkish
Arabic	-	LD (p = .004) MTLD (p = .003) LS1 (p = .001) LS2 (p = .002)	D (p = .02) MTLD (p < .001)	D (p = .01) MTLD (p < .001) LS1 (p = .002) LS2 (p < .001)	LD (p = .007) D (p < .001) MTLD (p = .002)	MTLD (p = .006)	LD (p < .001)	LD (p < .001) D (p < .001) MTLD (p < .001)	MTLD (p = .044)	LD (p < .001) D (p = .03) MTLD (p = .002)	D (p = .01) MTLD (p = .02)
Chinese		-	LS1 (p < .001) LS2 (p = .02)	LD (p = .02)	LS1 (p < .001) LS2 (p < .001)	LS1 (p = .045)		LS1 (p < .001)	LD (p = .02) LS1 (p = .001) LS2 (p = .046)	LS1 (p < .001) LS2 (p < .001)	LS1 (p = .001)
French			-	LS1 (p < .001) LS2 (p < .001)	LS2 (p = .02)		LD (p < .001)	LD (p < .001)		LD (p < .001) LS2 (p = .000)	
German				-	LD (p = .04) LS1 (p < .001) LS2 (p < .001)	LS2 (p = .02)	LD (p < .001) MTLD (p = .02)	LD (p < .001) LS1 (p = .001) LS2 (p < .001)	LS1 (p = .002) LS2 (p < .001)	LD (p < .001) LS1 (p < .001) LS2 (p < .001)	LS1 (p = .002) LS2 (p = .001)
Hindi					-	LS1 (p = .02) LS2 (p < .001)	D (p = .01) LS1 (p < .001) LS2 (p < .001)	LS2 (p < .001)	LD (p = .04) D (p = .01) LS2 (p = .002)		LS2 (p < .001)
Italian						-	LD (p = .001)	LD (p < .001)		LD (p = .001) LS1 (p = .009) LS2 (p < .001)	

Table 12 (Continued)

	Arabic	Chinese	French	German	Hindi	Italian	Japanese	Korean	Spanish	Telugu	Turkish
Japanese							-		LD (p < .001)	LS1 (p < .001) LS2 (p < .001)	
Korean								-	LD (p < .001) D (p = .03)	LS2 (p < .001)	LD (p = .01)
Spanish									-	LD (p < .001) LS2 (p = .001)	
Telugu										-	LS2 (p < .001)
Turkish											-

High Level

The significant differences in syntactic complexity between specific L1s in the high score level are displayed in Appendix B. Unlike the medium level, DC/C was the most prominent indicator of between-L1 differences as it distinguished the most L1 pairs (15), followed by MLS (12) and C/S, MLT, C/T, and CN/T (all 10 pairs). Again, Korean was included in a lot of comparison pairs (37). In the case of lexical complexity, while LD, LS1, and LS2 all distinguished over 10 L1 pairs, D and MTLD only mattered for the Hindi and Spanish pair. Hindi was also the most common L1 among the comparison pairs with 19 occurrences.

There were three accuracy measures that varied significantly across L1s in the high level but like the medium level, only the number of morphological errors emerged as a significant difference between L1s, specifically between Chinese and German (p = .04). Different from the low and medium score levels, W/Tx did not vary significantly across L1s in the high level.

Summary of CALF Variations Across L1s

To summarize, syntactic complexity measures varied to a great extent across L1s, with the largest number of significant between-L1 differences occurring in the medium score level. Korean and Telugu L2 writers had the most differences from other L1 groups. MLS, C/S, MLT, and CN/T stood out as the most common indicators of between-L1 differences across the score levels. Lexical complexity demonstrated both similar and different patterns of between-L1 variation. Like syntactic complexity, its largest number of differences was witnessed in the medium score level. However, Hindi was the L1 with the most differences from other L1s in terms of lexical complexity, and Turkish was the one with the fewest. In the medium level, lexical complexity measures separated Arabic from all other L1s. LD and LS2 were the most consistent among lexical complexity measures in separating many L1 pairs across the levels.

Accuracy and fluency measures varied across L1s to a lesser extent than syntactic and lexical complexity. For accuracy, there were only four significant between-L1 differences identified in all three score levels. The number of morphological errors was the only measure distinguishing between L1s in the all the levels. In terms of fluency, only 12 between-L1 differences were observed across levels.

RQ2: To What Extent do CALF Measures Vary Across Score Levels in Each L1 Group?

Table 13 presents the differences in CALF measures across three score levels in each of the 11 L1 groups. Among the 26 CALF measures, W/Tx (fluency) or the number of words written within a period of time was the only measure that varied significantly across score levels in every L1 ($p < .001$) with medium to large effect sizes. For most L1s, it was also the measure that observed a significant difference in every post hoc pairwise comparison (low-medium, medium-high, low-high). Other measures that also differentiated score levels in multiple L1s were the total number of errors (seven L1s) and lexical diversity measures: D (eight L1s) and MTLD (nine L1s). On the contrary, DC/C did not vary across score levels for any L1 backgrounds.

Table 13

Between-Score Level Differences in CALF in Each L1 Group

Measure	Arabic		Chinese		French		German		Hindi		Italian		Japanese		Korean		Spanish		Telugu		Turkish	
	p	ES	*p*	ES	*p*	ES	*p*	ES	*p*	ES	*p*	ES	*p*	ES	*p*	ES	*p*	ES	*p*	ES	*p*	ES
Syntactic complexity																						
MLS	.20	0.01	.30	0.02	**.02**	0.05	.20	0.02	**.01**	0.06	.80	<0.01	**.04**	0.05	**.006**	0.06	.40	0.01	**.02**	0.05	**.003**	0.07
T/S	.30	0.02	.20	0.03	.09	0.03	.50	0.01	**.005**	0.07	**.006**	0.07	.70	<0.01	.20	0.02	.20	0.02	.20	0.02	.09	0.03
C/S	.80	<0.01	.10	0.03	**.02**	0.04	.20	0.02	**.009**	0.06	.08	0.04	.60	-0.01	.40	-0.00	.60	<0.01	**.004**	0.06	**.001**	0.08
MLT	1	<0.01	.60	<0.01	.07	0.02	.30	0.02	.09	0.03	.10	0.03	**.02**	0.05	**.02**	0.04	1	<0.01	**.01**	0.04	**.04**	0.03
C/T	.90	<0.01	.10	0.03	**.03**	0.04	.09	0.03	**.03**	0.05	.40	0.01	.60	-0.01	.80	-0.01	.80	-0.01	**.003**	0.06	**.004**	0.07
CT/T	.30	<0.01	.64	-0.01	.70	-0.01	.25	0.01	.10	0.02	.06	0.03	.70	-0.01	.20	0.01	.85	-0.01	**.04**	0.03	.30	<0.01
DC/T	.60	<0.01	.20	0.02	**.04**	0.03	.20	0.02	.05	0.04	.30	0.02	.90	<0.01	.10	0.03	.90	<0.01	**.006**	0.05	**.006**	0.06
CP/T	.20	0.03	.30	0.02	.20	0.01	.40	0.01	.60	<0.01	**.001**	0.09	.07	0.04	**<.001**	0.11	.60	0.01	**.03**	0.04	.09	0.02
CN/T	.90	<0.01	.60	<0.01	.06	0.03	.20	0.02	.20	0.02	.10	0.03	**.02**	0.04	.30	0.02	.40	0.01	.06	0.03	**.04**	0.03
VP/T	1	<0.01	.10	0.03	**.03**	0.04	.30	0.02	**.03**	0.04	.30	0.02	.40	-0.00	.80	-0.01	1	<0.01	**.004**	0.06	**.005**	0.06
MLC	.30	<0.01	.10	0.03	.40	-0.00	**<.001**	0.10	.40	0.01	**.02**	0.04	**<.001**	0.14	**.002**	0.09	1	<0.01	**.005**	0.06	**<.001**	0.09
DC/C	.10	0.01	.42	-0.00	.40	-0.00	.40	0.01	.36	<0.01	.57	-0.00	.48	-0.00	.30	<0.01	.70	<0.01	.16	0.01	.05	0.03
CP/C	.70	-0.01	.09	0.02	.40	0.01	.06	0.04	**.04**	0.04	**<.001**	0.10	.10	0.03	**.001**	0.09	.70	<0.01	**.004**	0.07	.08	0.02
CN/C	.40	0.02	.40	0.01	.60	0.01	**.008**	0.06	.08	0.02	.11	0.02	**.01**	0.05	.60	<0.01	.40	0.40	.13	0.01	**.01**	0.05

Table 13 (Continued)

Measure	Arabic		Chinese		French		German		Hindi		Italian		Japanese		Korean		Spanish		Telugu		Turkish	
	p	ES	*p*	ES	*p*	ES	*p*	ES	*p*	ES	*p*	ES	*p*	ES	*p*	ES	*p*	ES	*p*	ES	*p*	ES
Lexical complexity																						
LD		-								<				-				-		-		-
	.45	0.00	**.02**	0.04	.40	0.01	**.003**	0.06	.29	0.01	**.003**	0.07	.40	0.00	**.03**	0.04	.76	0.01	.40	0.00	.56	0.00
D	<				<						<								<			<
	.001	0.12	**.02**	0.06	**.001**	0.10	.11	0.02	.06	0.03	**.001**	0.09	**.02**	0.04	**.03**	0.04	**.04**	0.04	**.001**	0.09	.35	0.01
MTLD	<								<		<								<			
	.001	0.17	**.006**	0.08	**.001**	0.09	.07	0.02	**.001**	0.13	**.001**	0.15	**.003**	0.07	**.002**	0.08	.19	0.01	**.001**	0.18	**.003**	0.07
LS1							<		<			<						-				
	.03	0.04	**.007**	0.07	.10	0.03	**.001**	0.13	**.001**	0.09	.21	0.01	.20	0.02	**.003**	0.07	.73	0.01	**.004**	0.05	.07	0.02
LS2						-	<					<		<				<				
	.10	0.02	**.001**	0.08	.61	0.01	**.001**	0.16	**.004**	0.06	.27	0.01	.21	0.01	**.03**	0.04	.30	0.01	**.03**	0.03	.10	0.03
Accuracy																						
Inc	.10	0.14	.08	0.18	**.01**	0.30	.30	0.03	**.02**	0.28	.07	0.18	**.003**	0.41	**.02**	0.28	.12	0.08	**.007**	0.29	.05	0.15
Morph	<							<										-				
	.001	0.53	.10	0.09	**.03**	0.17	.38	0.01	.05	0.20	.20	0.13	.40	0.07	.30	0.08	.41	0.00	**.04**	0.15	**.008**	0.24
Prep				-		-								<								-
	.02	0.19	.99	0.07	.92	0.07	.20	0.13	.50	0.05	**.03**	0.23	.37	0.01	.08	0.12	.27	0.02	**.02**	0.18	.52	0.02
Spell	**.04**	0.17	**.03**	0.18	.10	0.14	.26	0.03	.30	0.10	.20	0.13	.06	0.19	.20	0.10	.60	0.03	.09	0.10	**.04**	0.22
Synt	**.009**	0.26	**.03**	0.18	.40	0.06	**.003**	0.30	.16	0.06	.06	0.20	.30	0.08	**.01**	0.22	.20	0.20	.08	0.17	**.005**	0.37
Total	<															-					<	
	.001	0.47	**.02**	0.19	**.04**	0.16	.05	0.14	.08	0.11	**.04**	0.22	**.02**	0.19	.39	0.00	.40	0.34	**.009**	0.24	**.001**	0.47
Fluency																						
W/Tx	<		<		<		<		<		<		<		<		<		<		<	
	.001	0.32	**.001**	0.20	**.001**	0.25	**.001**	0.16	**.001**	0.08	**.001**	0.40	**.001**	0.41	**.001**	0.53	**.001**	0.33	**.001**	.18	**.001**	0.31

Note. Significant level $p < 0.05$. Significant *p* values are in bold. MLS = mean length of sentence, T/S = number of T-units per sentence, C/S = number of clauses per sentence, MLT = mean length of T-unit, C/T = number of clauses per T-unit, CT/T = complex T-unit ratio, DC/T = number of dependent clauses per T-unit, CP/T = number of coordinate phrases per T-unit, CN/T = number of complex nominals per T-unit, VP/T = number of verb phrases per T-unit, MLC = mean length of clause, DC/C = number of dependent clauses per clause, CP/C = number of coordinate phrases per clause, CN/C = number of complex nominals per clause. Inc = incomprehensible errors, morph = morphological errors, prep = preposition errors, spell = spelling errors, synt = syntactic errors, total = total number of errors. LD = lexical density, D = index of lexical diversity, MTLD = measure of textual lexical diversity, LS1 = lexical sophistication-I, LS2 = lexical sophistication-II. W/Tx = number of words per text.

Overall, Korean, Turkish, and Telugu registered the highest numbers of CALF measures that distinguished score levels, 13, 15, and 19 (of 26), respectively. The other L1s have from seven to 11 measures, except for Spanish, which has only two measures that differentiated score levels: D and W/Tx.

Arabic

When each L1 is considered, Arabic has 10 measures that varied significantly across score levels, none of which were syntactic complexity measures. Five were accuracy measures (all except incomprehensible errors), which were all able to separate low and high levels of writing quality. Among the other significant measures (D, MTLD, LS1, LS2, and W/Tx), only MTLD and W/Tx were able to separate all three levels.

Chinese

Chinese has nine measures that varied significantly across score levels, which are LD, D, MTLD, LS1, LS2, spelling, syntactic, and total errors, and W/Tx. Post hoc tests showed significant differences between medium and high score levels for LD, LS1, LS2, and W/Tx, between low and high for D, MTLD, spelling and total errors, and W/Tx, and between low and medium for syntactic errors and W/Tx,

French

Eleven of 26 CALF measures varied significantly across score levels for French L2 writers. There are five significant syntactic complexity measures – MLS, C/S, C/T, DC/T, and VP/T. However, they were not able to distinguish specific score levels. D, MTLD, syntactic errors, and W/Tx differed significantly between low and high levels. D, MTLD, and W/Tx also differed significantly between low and medium levels. Morphological errors and W/Tx differed significantly between medium and high levels.

German

German only has seven measures that differed significantly across score levels: MLC, CN/C, LD, LS1, LS2, syntactic errors, and W/Tx. Among these measures, MLC, CN/C, and W/Tx differentiated low from medium and high levels. LD, LS1, and LS2 distinguished between medium and high levels, whereas syntactic errors and W/Tx distinguished between low and high levels.

Hindi

Hindi L2 writers have eleven CALF measures that differentiated score levels like French L2 writers. However, pairwise comparisons revealed that MLS, C/T, and VP/T were not significantly different between any score levels, while T/S, C/S, and CP/C were significantly different between medium and high levels. In addition, LS2 separated medium and high levels, and MTLD, LS1, and W/Tx differentiated high from low and medium levels.

Italian

Ten CALF measures of Italian L2 essays varied significantly across score levels, including T/S, CP/T, MLC, CP/C, LD, D, MTLD, preposition and total errors, and W/Tx. Most of these measures could distinguish between low-high and/or medium-high levels (T/S, CP/T, MLC, CP/C, LD, D, MTLD, W/Tx). MTLD and W/Tx were the only measures that could differentiate low from medium levels of scores.

Japanese

Similarly, ten measures separated score levels for Japanese L2 writers: MLS, MLT, CN/T, MLC, CN/C, D, MTLD, incomprehensible errors, total errors, and W/Tx. All but MLS and incomprehensible errors differentiated high from low and/or medium levels. Only W/Tx differentiated between low and medium levels.

Korean

In the case of Korean, 13 measures were found to vary significantly across score levels, five of which were syntactic complexity measures. Nine of these measures (MLS, MLT, CP/T, MLC, CP/C, D, MTLD, LS1, and W/Tx) were significantly different between low and high levels, five (MLC, LS1, LS2, syntactic errors, and W/Tx) were significantly different between medium and high levels, and five (MLS, CP/T, LD, MTLD, and W/Tx) were significantly differently between low and medium levels.

Spanish

As mentioned above, nearly all CALF measures (and all syntactic complexity and accuracy measures) failed to separate the score levels of Spanish L2 writers, except for D and W/Tx, with the latter differentiating all three levels.

Telugu

Of the 11 L1 groups, Telugu L2 writers had the most CALF measures varying significantly across score levels, 10 of which were syntactic complexity measures. The syntactic complexity measures that were not significantly different across levels were T/S, CN/T, DC/C, and CN/C. The significant measures varied between medium and high levels (MLS, C/S, MLT, C/T, CT/T, DC/T, VP/T), low and medium levels (CP/T), and low and high levels (CP/C). LS1 and LS2 varied between medium and high levels, whereas incompressible, morphological, preposition, and total errors varied between low and medium and/or low and high levels. D, MTLD, and W/Tx were the three measures that differentiated all three score levels.

Turkish

Lastly, 15 of the measures emerged significantly different across the score levels of Turkish L2 writers. Many of these measures differentiated between medium and high levels

(MLS, C/S, C/T, DC/T, MLC, CN/C, MTLD, morphological, syntactic, and total errors, and W/Tx). MLC and morphological, syntactic, and total errors also differentiated between low and high levels. Only W/Tx differentiated all three score levels.

Summary of CALF Measures Across Score Levels

In summary, most L1s have measures related to lexical complexity (e.g., D, MTLD), accuracy (e.g., total errors), and fluency (i.e., W/Tx) that varied significantly across score levels. Additionally, most L1s have very few syntactic complexity measures that differentiated levels. French, German, Hindi, Italian, Japanese, and Korean have from two to six measures that differentiated score levels, while Arabic, Chinese, and Spanish have none. In contrast, Telugu and Turkish L2 writers have many significant syntactic complexity measures, making them the L1s with the most measures varying significantly across score levels. Overall, it seems that the measures that varied significantly across score levels varied by language, and there was no consistent measure or set of measures that differentiated score levels across all languages, except for W/Tx.

RQ3: To What Extent Can CALF Measures and L1 Backgrounds Predict Score Levels?

Before multinominal logistic regression models were fit to discover whether CALF measures and L1 backgrounds are predictive of different L2 writing score levels, variable selection was performed to satisfy the assumption of no multicollinearity. Multicollinearity occurs when two or more independent variables or predictors are highly correlated with each other. Being selective with the predictors also simplifies a model, making it easier to interpret. For these reasons, one-way ANOVA analyses with CALF measures as the dependent variables and writing quality as the independent variable were conducted to identify potential predictors that could be included in the models. The measures MLS, T/S, C/S, MLT, C/T, CT/T, DC/T,

VP/T, MLC, CP/C, CN/C, LD, D, MTLD, LS1, LS2, and W/Tx were found to differ significantly between score levels. However, C/S, MLT, C/T, CT/T, DC/T, VP/T, D, and LS2 were excluded because they were multicollinear with the other measures ($r > .70$) while demonstrating smaller effect sizes in the relationship with writing quality than those measures.

In the end, nine CALF measures were entered in the first multinominal logistic regression model: MLS, T/S, MLC, CP/C, CN/C, LD, MTLD, LS1, and W/Tx, along with L1 (11 levels). With McFadden's $R^2 = 0.32$, the model fits well (Levshina, 2015). Table 14 presents information about the accuracy of the model (i.e., how correctly the model predicts the score levels of essays). The rows display the number of observations (i.e., essays) in each score level, and the columns show how many essays would be predicted to be in a score level. The diagonal (from left to right) shows how many essays observed in a level would be predicted to be in that level. Using the low score level as an example, there is a total of 198 essays (99 + 97 + 2) observed in this level. The model would predict 99 of these low score essays as low, 97 as medium, and two as high. The accuracy of the model is 0.70, which can be calculated by adding the numbers of correct predictions in the diagonal (99 + 807 + 279) and dividing the total (1,185) by the number of observations (1,683).

Table 14

Accuracy of the First Multinomial Logistic Regression Model

	Low	Medium	High
Low	99	97	2
Medium	18	807	127
High	0	254	279

In addition to the accuracy measure, the model also reported measures of recall and precision (see Table 15). The measure recall shows the proportion of essays of each score level predicted by the algorithm. The results indicate that 85% of medium-score essays would be predicted as medium-score essays by the model. Low- and high-score essays, unfortunately, have relatively low recall values, about 50%. The measure precision, in contrast, shows how many times the predictions of score levels made by the model were correct. The low score level was predicted most accurately (85%), whereas the precision value was lower for the medium level (70%) and the high level (68%).

Table 15

Recall and Precision of the First Multinomial Logistic Regression Model

	Low	Medium	High
Recall	0.50	0.85	0.52
Precision	0.85	0.70	0.68

Note. All values represent percentages.

Table 16 shows the log odds ratios of the CALF measures and L1 backgrounds. It indicates that high-score writers used significantly less MLS ($p = .007$) but more CP/C ($p = .003$). They also had significantly higher LD ($p = .03$), MTLD ($p < .001$), LS1 ($p < .001$), and W/Tx ($p < .001$). They are also likely to be German ($p = .002$). On the contrary, low-score writers demonstrated significantly more MLS ($p = .02$) and lower MTLD ($p < .001$) and W/Tx ($p < .001$).

Table 16

Coefficients and P-Values of the First Multinomial Logistic Regression Model

	Low		Medium		High	
	Log-odds	*p*	Log-odds	*p*	Log-odds	*p*
Intercept	11.09	< **.001**	2.70	**.001**	-10.99	< **.001**
MLS	0.008	**.02**	0.001	.65	-0.02	**.007**
T/S	0.32	.36	0.58	**.006**	-0.39	.26
MLC	-0.16	.21	0.02	.42	-0.004	.91
CP/C	-0.53	.62	-0.64	.16	1.72	**.003**
CN/C	0.18	.78	0.18	.43	-0.16	.60
LD	0.39	.90	-5.06	**.003**	4.86	**.03**
MTLD	-0.08	< **.001**	-0.003	.52	0.04	< **.001**
LS1	5.96	**.002**	-6.33	< **.001**	5.14	< **.001**
W/Tx	-0.03	< **.001**	-0.001	.18	0.02	< **.001**
Chinese	0.19	.68	1.00	< **.001**	-1.55	< **.001**
French	-1.16	**.03**	0.31	.21	0.19	.55
German	-1.20	.10	-0.51	.05	0.98	**.002**
Hindi	0.02	.97	0.30	.23	-0.13	.67
Italian	-1.19	**.002**	0.34	.19	0.12	.73
Japanese	-0.49	.29	0.72	**.006**	-0.68	.05
Korean	0.02	.97	0.41	.12	-0.75	**.03**
Spanish	-0.31	.50	0.08	.75	0.34	.26
Telugu	-0.002	.10	0.81	**.001**	-0.59	.07
Turkish	-1.51	**.004**	0.51	**.04**	0.009	.98

Note. Significant level $p < 0.05$. Significant *p* values are in bold. MLS = mean length of sentence, T/S = number of T-units per sentence, MLC = mean length of clause, CP/C = number of coordinate phrases per clause, CN/C = number of complex nominals per clause. LD = lexical density, MTLD = measure of textual lexical diversity, LS1 = lexical sophistication-I. W/Tx = number of words per text.

The same variable pruning procedure was applied for the second multinomial logistic regression model for the accuracy subset. A second model was run because accuracy was measured based on 329 essays rather than the full dataset of 1,683 essays as in the case of syntactic complexity, lexical complexity, and fluency. Three accuracy measures remained in the model: incomprehensible errors, preposition errors, and the total number of errors. The output of the model reported a McFadden statistic of 0.16, indicating a poor fit. The accuracy of the model is reported as 0.54 (see Table 17). Table 18 reports the recall and precision estimates of the model. Since the model did not fit well, the log odds ratios of the predictors are not reported.

Table 17

Accuracy of the Second Multinomial Logistic Regression Model

	Low	Medium	High
Low	60	15	26
Medium	34	26	54
High	7	16	91

Table 18

Recall and Precision of the Second Multinomial Logistic Regression Model

	Low	Medium	High
Recall	0.59	0.23	0.80
Precision	0.59	0.46	0.53

Note. All values represent percentages.

CHAPTER FIVE: DISCUSSION

Relationship Between CALF and L1 Backgrounds

Although there have been many studies examining differences in CALF between L1 and L2 writers (e.g., Ai & Lu, 2013; Crossley & McNamara, 2009; Eckstein & Ferris, 2018; Lu & Ai, 2015), few have explored differences among L2 writers of different L1 backgrounds. The first research question of this study thus asked whether CALF measures of EFL writers differed based on their L1 backgrounds in three separate score levels: low, medium, and high. One-way ANOVAs and their equivalents showed that in each score level, many CALF measures varied significantly across L1s. First, all syntactic complexity measures varied significantly across L1s in the medium and high levels. In the low level, there were significant differences in all the measures except for CT/T, CP/C, and CN/C. In all three levels, significant between-L1 differences were found in all five dimensions of syntactic complexity: length of production units (MLS, MLT, MLC), amount of subordination (C/T, CT/T, DC/T, DC/C), amount of coordination (CP/T, T/S, CP/C), degree of phrasal sophistication (CN/T, VP/T, CN/C), and overall sentence complexity (C/S).

Similar results have been reported in previous research. Studying syntactic complexity in college-level English writing, Lu and Ai (2015) discovered significant differences between eight L1 groups (Bulgarian, Chinese, English, French, German, Japanese, Russian, and Tswana) in all 14 SCA measures. Moreover, they found that only three of the measures differed significantly between the NS group and the NNS groups when the latter was treated as a single entity. Mancilla et al. (2017) also found significant differences in only four out of 10 syntactic

complexity measures between NSs and a heterogenous group of NNS writers in asynchronous online discussions. On the contrary, Ai and Lu (2013) found eight out of the same 10 measures to differ significantly between NSs and a homogenous NNS group (i.e., Chinese learners of English). The difference in the number of syntactic complexity measures emerging as significant when comparing NS and homogeneous/heterogenous NNS groups of writers together with the present findings that syntactic complexity generally distinguishes between individual L1 groups suggest that treating NNS writers of heterogeneous L1 backgrounds as one group may obscure L1-related differences in syntactic complexity. They also lend support to previous claims that heterogenous L1 groups should be viewed independently in L2 writing research (Lu & Ai, 2015; Ortega, 2015).

As for why such between-group differences exist for syntactic complexity in L2 writing, it could potentially be attributed to L1 influences. It has been speculated and demonstrated in the L2 writing literature that L1 transfer, either positive or negative, has an impact on various structural aspects of L2 writing (Jarvis & Crossley, 2012; Liu, 2008; Rankin, 2012; Uysal, 2008; van Vuuren, 2013; van Weijen et al., 2009). Given that other factors such as topic and writing quality were accounted for in the current study, there is even more reason to think that learners' L1 can decide the syntactic complexity manifested in their L2 writing. However, the sizes of the differences across L1 groups observed in this study varied to a great extent in all score levels. This means although learners' L1 may account for a certain portion of the variation in syntactic complexity in L2 writing, it cannot account for all the variation. The rest of this variation, as shown by previous research, may derive from a number of factors, such as development and instruction (Mazgutova & Kormos, 2015; Vyatkina et al., 2015) and cognitive task complexity (Johnson, 2017).

It should also be noted that the breadth of the differences was also limited to certain L1 pairs. The fact that the 11 languages examined can be classified into seven language groups could have limited the impact L1 has on syntactic complexity measures. Vo and Barrot (2022) argued that typologically similar languages could share remarkable similarities in CALF measures. Their results, for example, showed that Chinese and Thai L2 writers had striking similarities in MLS, MLT, T/S, and CN/T. Such evidence can be found in the current study. For instance, French, Italian, and Spanish (which belong to the Romance language family) L2 writers were barely significantly different from each other in syntactic complexity across all score levels. The same is true for Japanese, Korean, and Turkish (which belong to the Altaic language family) L2 writers. At the same time, L2 writers from typologically different language backgrounds such as Arabic and Korean differed significantly from one another in various syntactic complexity measures. These results indicate that the extent to which syntactic complexity measures vary across L1 groups may depend on how the L1s are related to each other.

For lexical complexity, all measures varied significantly as a function of L1 in the medium and high score levels, whereas there were no significant differences in D and LS1 across L1s in the low score level. These results resemble the patterns observed for syntactic complexity above. For both syntactic and lexical complexity, there were fewer measures that emerged as significant separators of L1s in the low level compared with the other two levels. It could be that on this level where L2 writers just started, they have not learned enough syntactical structures and vocabulary for certain syntactic and lexical complexity measures to be reflected in their writing.

Like investigations into differences in syntactic complexity among L2 writers, research on the relationship between lexical complexity and L1 is still scarce, despite many studies

comparing lexical complexity in L1 and L2 writing (e.g., Crossley & McNamara, 2009; Eckstein & Ferris, 2017; Eckstein & Chang, 2022; Rahayu et al., 2021). Nevertheless, the limited work in this area seems to be in line with the significant effects of L1 on lexical complexity reported in this study. Jarvis (2002), for example, found significant differences in lexical diversity (as measured by an algebraic transformation of TTR – the Uber index) between Finnish and Swedish L2 writers in their English narrative writing. Furthermore, Crossley and McNamara (2012a) reported significant differences in lexical diversity (as measured by MTLD) between all L1 pairs except Spanish and Czech when examining English essays written by L1 Czech, Finnish, German, and Spanish speakers.

In terms of accuracy, the number of morphological errors and the total number errors (per 100 words) seem to be consistent separators of L1s as they varied significantly across L1s in all three score levels. In the medium level, there were also significant between-group differences for preposition errors. From a typology standpoint, it would make the most sense for morphological and syntactic errors to be able to differentiate L2 writers of one L1 background from others because morphology and syntax are two main areas of linguistic typology and writing accuracy may increase if there are typological similarities between the L1 and the target language and between L1s themselves (i.e., positive transfer). However, only morphological errors support this hypothesis by successfully distinguishing between typologically different L1s such as German and Korean. Syntactic errors, on the other hand, were just on the verge of significance with *p*-values of .06 in the low level and .05 in the high level.

Barrot and Gabinete (2019) also argued for the potential role of the ESL vs. EFL status in producing accurate English essays after finding that Filipino and Singaporean ESL learners wrote more accurate essays than a heterogenous group consisting of Chinese, Indonesian,

Japanese, Korean, Taiwanese, and Thai EFL learners. However, this may not be the case in the current study as the Hindi and Telugu L1 groups, who belong to the "Outer Circle" of World Englishes (Kachru, 1985), did not demonstrate a significantly lower number of errors than several other groups that are in the "Expanding Circle" or are not in any circles.

Finally, fluency or the number of words per text varied significantly across L1s in the low and medium score levels. It was not able to separate the L1s in the high level. This may be because high-level L2 writers are aware of the minimum length recommended for a successful essay submitted as part of a standardized test and, at the same time, know not to go off limits but communicate ideas in succinct ways with their linguistic abilities. A further look into the differences in the low and medium levels revealed that Hindi and Telugu L2 writers tended to write longer essays than those from other L1 backgrounds. This is interesting because in theory, writers whose L1 is syntactically similar to English might be assumed to be more fluent in their English writing than writers whose L1 is syntactically different (Ringbom, 2007; Ringbom & Jarvis, 2009). However, the sentence structure of Hindi and Telugu (subject-object-verb) generally differs from that of English (subject-verb-object), and writers from those L1 backgrounds were the most fluent groups in the current study. This might be because Hindi and Telugu L2 writers oftentimes use English as an L2 in India, and therefore, may be more comfortable with expressing their thoughts in writing.

Relationship Between CALF and L2 Writing Quality

The second research question investigated whether CALF measures differed significantly based on score levels, which are equated with writing quality, for each L1 group of L2 writers. Results from statistical tests indicated that the relationship between CALF measures and L2 writing quality, operationalized as score levels, could be significant, depending on the specific

measures used and the L1 of the writers. Syntactic complexity measures seem not to be very consistent indicators of writing quality when their most consistent indicators, MLS and MLC, could only separate score levels across six L1 backgrounds. Moreover, no syntactic complexity measures could distinguish between the score levels of the Arabic, Chinese, and Spanish L2 groups, while only two were able to differentiate the levels of the German group. These findings are somewhat surprising when a positive correlation between syntactic complexity and L2 writing quality has been consistently reported in previous studies (e.g., Barrot & Agdeppa, 2021; Kim & Crossley, 2018; Zhang & Lu, 2022). Nevertheless, the findings that length of production unit measures, including MLT (which differentiated score levels across three L1s), were significant indicators of writing quality are partially aligned with previous studies reporting that MLS, MLT, and/or MLC are correlated with proficiency levels and subjective ratings of writing quality (Bulté & Housen, 2014; Gyllstad et al., 2014; Khushik & Huhta, 2019; Lu, 2011; Vo & Barrot, 2022).

Therefore, the relationship between syntactic complexity and L2 writing quality may not be as straightforward as it seems. In some cases, excessive complexity could hinder comprehension and lead to errors or confusion. Particularly in high-stakes testing situations, learners are less likely to take risks and thus prioritize being accurate, while raters are also sensitive to writers' accuracy (Fritz & Ruegg, 2013; Kim & Kessler, 2022). L2 writers have also been shown to prioritize accuracy over syntactic complexity when dealing with unfamiliar topics (Kessler et al., 2022). Such contextual prioritization is partly supported by the current study's finding that the total number of errors was one of the most consistent separators of score levels. Overall, rather than trying to use more complex syntactic structures, it may be more important

for L2 learners to use those structures appropriately and accurately to enhance the overall quality of their writing.

Regarding lexical complexity measures, they tend to be more consistent separators of score levels compared to syntactic complexity and accuracy measures. D and MTLD were able to differentiate score levels across eight and nine L1s, respectively. These measures plus LD could differentiate high from low and medium levels well. This may be because lexical complexity allows writers to convey their ideas more precisely and to express a range of nuanced meanings. It also adds depth and richness to the writing, making it more engaging and interesting to read. These findings resonate with Bulté and Housen's (2014) that lexical richness, determined by G – a measure related to D, is a robust indicator of higher writing quality.

In terms of accuracy, the total number of errors per 100 words was the most consistent separator of score levels across L1s (especially between low-high levels). As expected, fewer errors are more likely to lead to more understanding and thus higher evaluations of writing. The ability of the total number of errors to separate score levels may also be attributed to the reported reliability and validity of the constituent morphological, preposition, spelling, and syntactic errors examined in this study (Yoon & Polio, 2017). Unlike Verspoor et al. (2012), who found the number of spelling errors to distinguish between two adjacent CEFR levels of A1.2 and A2.1, I found it to discriminate between low and high levels for two L1s where its effects on score levels were significant. A closer look at the results revealed that accuracy measures mostly differentiated low-high levels across L1s rather than adjacent levels of low-medium and medium-high.

Fluency was simply and solely operationalized as text length or the total of number of words per essay (W/Tx), which showed a generally linear progression with score levels. W/Tx

was also the most consistent indicator of writing quality in this study as it could distinguish score levels regardless of L1. This finding appears to disagree with Vo and Barrot's (2022) observation that the number of words per text could differentiate proficiency levels in only two L1 groups (Chinese and Korean). However, there are potentially important distinctions to make between the two studies. The current study did not include L1 backgrounds that were examined in Vo and Barrot such as Indonesian, Pakistani, and Thai, so the effects of fluency on writing quality for these L1s are unclear. In addition, the focus of the latter was language proficiency, not precisely writing quality. Overall, the mostly linear progression in text length across score levels found in this study corroborates previous findings that essays of higher proficiency levels tend to be longer (De Angelis & Jessner, 2012; Vo & Barrot, 2022).

Predictive Power of CALF and L1 on L2 Writing Quality

The purpose of the third research question was to test the predictive power of CALF measures and L1 on L2 writing quality. To answer this question, I fitted two multinominal regression models after careful selection of variables: one with complexity and fluency measures using the full dataset (1,683 essays) and one with accuracy measures using the hand-coded subset (329 essays). The first model had MLS, T/S, MLC, CP/C, CN/C, LD, MTLD, LS1, and W/Tx as predictors. In the second model, incomprehensible errors, preposition errors, and total errors were included as predictors. L1 was also added to both models as a predictor. The results of the multinominal regression analysis showed that with each increase in CP/C, LD, MTLD, LS1 and W/Tx, the chances of an essay being of high quality, as indicated by the positive log odds, were significantly higher by 1.72, 4.86, 0.04, 5.14, and 0.02, respectively. In other words, compared to other essays, high score essays contained significantly more CP/C and were significantly longer and more lexically diverse and sophisticated.

The finding regarding CP/C in this study is strongly supported by previous research. Coordinate phrases have been found to be not only significantly correlated with but also predictive of score/proficiency levels (Barrot & Agdeppa, 2021; Kim, 2014; Kim & Crossley, 2018; Lu, 2010, 2011; Zhang & Lu, 2022). It is likely that L2 learners who have a higher CP/C ratio have a better grasp of the different types of phrases and conjunctions of the target language and are more able to produce more complex sentences with greater accuracy. An understanding of coordinate phrases can bring learners to a new level because these phrases usually constitute basic parts of a sentence like the subject and object(s) while carrying additional information. They can also dictate verb forms when functioning as the subject. In contrast, learners with a lower CP/C ratio may struggle with complex sentence structures, leading to poor quality writing that is less accurate and less cohesive.

Lexical density, diversity, and sophistication are generally associated with more advanced writing skills, as demonstrated by LD, MTLD, and LS1 being indicative of high writing scores. When a writer uses more sophisticated content words and many different words, they are better able to express themselves clearly and precisely, convey more complex ideas, and provide more nuanced descriptions, leading to higher writing quality. Additionally, MTLD takes into account the length of the text and performs well in terms of validity (McCarthy & Jarvis, 2010). It is stable in short L2 texts (Zenker & Kyle, 2021), which is true in the case of the current dataset. All these factors might have contributed to the predictive power of the measure.

Although it could be a byproduct of increased syntactic and lexical complexity, fluency or text length was not correlated with any of the complexity measures in the study. Hence, its ability to predict score levels can be explained with respect to content. Higher-scored writers might have developed more ideas and arguments to support their opinion. This is particularly

true for the TOEFL Independent Writing Task that appreciates well-developed essays with appropriate explanations exemplifications and/or details. In other studies, longer texts were also shown to distinguish and predict writing quality (Ai & Lu, 2013; Grant & Ginther, 2000; Kim, 2014; McNamara et al., 2010).

While syntactic complexity, lexical complexity, and fluency all demonstrated some predictive power on the score levels of test-takers, accuracy failed to do so. Theoretically, linguistic accuracy can predict L2 writing quality because when a writer is accurate in their use of language, they are more likely to produce texts that are easier to read and understand and are less likely to be distracting to the reader. Accuracy is particularly important when the writer does not have a full command of the target language and struggles to produce grammatically correct sentences. Inaccuracies can result in confusion, miscommunication, and an overall negative impression of the writer's ability. In reality, this may not be the case as raters may ignore certain types of errors. For example, when coding the accuracy subset, I noticed that a lot of errors were repeated. If those repeated errors had been marked negatively every time they were encountered, there would be a lot more low score essays in the corpus. Thus, the lack of predictive power of accuracy on L2 writing quality may be contextual. It could also be that there was not sufficient data for any significant relationships to be uncovered. The regression model for accuracy measures was fitted on 329 essays, which is five times less than the input for the complexity and fluency model.

Overall, the multinominal regression models indicated that a certain dimension of CALF alone cannot account for L2 writing quality. The co-occurrence of measures representing different dimensions of CALF as predictors of high-quality writing in the first model demonstrated that different factors must be considered for a comprehensive and effective

assessment of L2 writing, providing empirical support for the argument that CALF is multifaceted and should be measured as such (Norris & Ortega, 2009).

CHAPTER SIX: CONCLUSION

Summary of Findings

This dissertation investigated the extent to which CALF measures in the essays of EFL writers varied across 11 L1 backgrounds and three levels of writing quality as well as the ability of CALF measures and L1 to predict writing quality. Overall, the results showed that 26 CALF measures differed at varying degrees as a function of L1 and writing quality. Most CALF measures varied significantly based on L1 with the effects being more pronounced in the medium and high score levels. Many CALF measures also varied significantly based on score levels, but the effects of score levels on CALF were dependent upon L1. This means that CALF measures found to separate score levels in one L1 may not have the same impact in another. Moreover, D, MTLD, the total number of errors, and W/Tx were the most consistent measures in distinguishing score levels. CP/C, LD, MTLD, LS1, and W/Tx together were also predictive of high scores. The findings of this study help build the foundation for a more thorough understanding of the role L1 plays in L2 writing as well as the importance of CALF to L2 writing quality. They have important implications for L2 writing assessment, pedagogy, and research.

Implications for L2 Writing Assessment

CALF measures are generally believed and have been demonstrated to be positively related to L2 writing quality. Therefore, measuring CALF can be a useful tool in assessing L2 writing quality. However, as shown in this study, CALF measures do not differentiate writing quality the same way across L1 backgrounds, suggesting that a "one-size-fits-all" approach to

assessing L2 writing may not be appropriate. Instead, L2 writing assessment should take into account the L1 background of the writer and consider, for example, the linguistic complexity of their L1 as a factor when evaluating the complexity of their L2 writing. In other words, CALF measures as indices of L2 writing quality may need to be adjusted to each L1 background or groups of L1 backgrounds. This would allow for a more nuanced approach to assessing L2 writing and contribute to fairness and equity. Otherwise, empirically powerful measures such as CP/C, MTLD, and W/Tx may be used. It is also recommended to use multiple CALF measures together to accurately assess L2 writing quality. Additionally, it is important to consider the writing task and genre, as different tasks and genres may require different levels of CALF, especially syntactic complexity.

Implications for L2 Writing Pedagogy

The findings of this study also have implications for both L2 writing instruction in general and test preparation in particular. Demonstrating that most of CALF measures vary significantly across L1s and different CALF measures vary significantly across score levels for different L1s, the study generally suggests teachers should be cautious of potential cultural influences and transfer issues in L2 writing. Differences in CALF between L1s may be meaningful if one L1 group performs better than another in a heterogeneous classroom. Within an L1 group, differences may occur between students of different levels. Investigating these areas of differences may help teachers adjust their teaching focus and methods to achieve the best results.

Among the CALF measurement tools used in the current study, TAASSC, TAALED, and LCA are quite easy to use, although it may take a little training. L2 writing teachers might use those tools to gain insights into the CALF of students' writing. For example, teachers may find

that syntactic complexity does not separate between low- and high-level L1 Arabic students, but accuracy does, like in the current study. They could then spend more time on the accuracy aspect of their students' writing. In the future, there may be better analysis tools that are optimized for L2 writing pedagogy.

For test preparation purposes, teachers should make it clear to all learners at what length they should write or at least help them set a minimum length as the findings showed that text length was the most consistent separators of score levels across L1s and that high-level test-takers tended to write longer. With D and MTLD also being relatively consistent in separating score levels across L1s, heterogenous L1 groups of learners may benefit from focusing on diversifying their vocabulary. Possible instructional methods include examining sample texts for lexical diversity and verbalizing the metacognitive process of word selection and/or revision during writing (González, 2017). Feedback on the length of sentences, the use of coordinate phrases, and the sophistication of vocabulary may also be valuable in test preparation courses as shorter sentences, more CP/C, and more sophisticated vocabulary were demonstrated to be indicative of highly scored essays in this study, in addition to longer texts and more diverse vocabulary. Given the usually short time frame of test preparation courses, teachers may find it more efficient to help students score higher by incorporating tasks that promote these features. For example, teachers might develop whole class, focused lessons and ask students to identify, review, and revise coordinate phrases in their practice texts using both teacher and peer feedback. In particular, teachers may need to pay more attention to Chinese and Korean L1 learners who are preparing for the TOEFL because they were shown to be less likely to achieve the high level of writing compared to other L1 groups in this study. Depending on the learning goals, L2

writing teachers may want to tailor their instruction to meet the unique needs of and address the challenges faced by L2 learners from various L1 backgrounds.

Implications for L2 Writing Research

This study has several important implications for L2 writing research. First, it demonstrated that a significant portion of writing quality can be explained by CALF, as shown by the regression model. This does not mean that CALF measures alone should be used to assess writing quality. However, when evaluating writing based on CALF, different CALF dimensions must be taken into consideration. Also, certain CALF measures seem to be more useful than others in terms of the ability to reflect writing quality. Researchers might thus consider using them in their future studies, starting with the measures that contributed to the regression model in the current study (i.e., MLS, T/S, MLC, CP/C, CN/C, LD, MTLD, LS1, and W/Tx). It may be worth investigating how pedagogical interventions involving these measures can help learners achieve higher scores in standardized tests.

Furthermore, the findings confirm the need for L2 writing scholars to treat L1 as an independent variable and to not cluster heterogenous L1 writers in one group. This would remove the possibility of L1 influence on the variability of CALF and reduce the variation in CALF in L2 writing. As can be observed in the literature, CALF studies, which mostly do not separate L2 learners from different L1 backgrounds, have reported inconsistent results and have had difficulties reaching a consensus regarding the relationship between CALF and writing quality. Controlling for L1 or taking it into account, therefore, would move the field forward by giving more specific results and enable comparisons between studies.

For L2 writing instructors, more L1-specific implications from research may be provided. In the area of NLI, where the goal is to accurately identify a writer's L1 based on their writing in

an L2, L1-specific findings are likely to help improve the accuracy of NLI systems, which can in turn be useful for a variety of applications in many areas, including language assessment, forensic linguistics, and machine translation, among others.

Limitations and Directions for Future Research

While this study attempted to tackle the issue of L1 in regard to CALF and L2 writing quality using all the resources that were available, it undoubtedly has several limitations that should be acknowledged. One limitation is that I did not examine fine-grained CALF measures. Fine-grained measures allow for a more nuanced assessment of L2 writing and facilitate more accurate comparisons across different L2 writing contexts and populations, which could have been important given the heterogeneity of L2 writers in the study. Large-grained measures, as the ones used in the current study, have an issue of granularity (Larsen-Freeman, 2009; Norris & Ortega, 2009; Wolfe-Quintero et al., 1998). They do not provide information about the constituent structures. For instance, the measure of MLT does not tell us which specific structures contribute to the length of the T-units observed. However, given the already large number of measures examined in the current study, fine-grained measures were not included.

Second, this study, like most studies in this research area, relied on quantitative measures, which may not provide a complete picture of the complexities of L2 writing. A qualitative look into the texts may provide insightful information on how certain L1s stand out, as in the cases of Korean, Telugu, and Turkish in this study. Future researchers could adopt mixed methods with a qualitative case study component to provide a rich and detailed understanding of the relationships between CALF and L2 writing quality by examining the use of language in context and uncovering the strategies that writers use to produce successful written texts. A case study may involve observing a single writer over an extended period, collecting data on the writer's

writing process, and analyzing the texts produced. Through this approach, the researcher may be able to identify specific strategies that the writer uses to produce complex, accurate, and fluent text, as well as areas where the writer struggles.

Another limitation is that the current dataset is limited to argumentative essays in a testing environment. More studies are needed on other genres and writing contexts to compare and evaluate the generalizability of the results. Finally, investigating the interactions between L1 and other factors may be worthwhile as L1 alone cannot fully account for the differences in CALF measures, as was indicated in the current dissertation. Topic familiarity, for example, has been shown to affect both text quality and CALF in L2 writing (Kessler et al., 2022; Yoon, 2017). However, data on such factors are not always available to researchers. Future research, if possible, could attempt to account for these factors and further examine their relationship to linguistic features and L1 in L2 writing.

REFERENCES

Abdel Latif, M. M. M. (2013). What do we mean by writing fluency and how can it be validly measured?. *Applied Linguistics*, *34*(1), 99–105. https://doi.org/10.1093/applin/ams073

Ädel, A., & Erman, B. (2012). Recurrent word combinations in academic writing by native and non-native speakers of English: A lexical bundles approach. *English for Specific Purposes*, *31*(2), 81–92. https://doi.org/10.1016/j.esp.2011.08.004

Ai, H., & Lu, X. (2013). A corpus-based comparison of syntactic complexity in NNS and NS university students' writing. In A. Díaz-Negrillo, N. Ballier, & P. Thompson (Eds.), *Studies in corpus linguistics* (pp. 249–264). John Benjamins Publishing Company.

Amiryousefi, M. (2016). The differential effects of two types of task repetition on the complexity, accuracy, and fluency in computer-mediated L2 written production: A focus on computer anxiety. *Computer Assisted Language Learning*, *29*(5), 1052–1068. https://doi.org/10.1080/09588221.2016.1170040

Ansarifar, A., Shahriari, H., & Pishghadam, R. (2018). Phrasal complexity in academic writing: A comparison of abstracts written by graduate students and expert writers in applied linguistics. *Journal of English for Academic Purposes*, *31*, 58–71. https://doi.org/10.1016/j.jeap.2017.12.008

Arppe A (2013). polytomous: Polytomous logistic regression for fixed and mixed effects. R package version 0.1.6. https://CRAN.R-project.org/package=polytomous.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press.

Barrot, J. S., & Agdeppa, J. Y. (2021). Complexity, accuracy, and fluency as indices of college-level L2 writers' proficiency. *Assessing Writing*, *47*, 1–11. https://doi.org /10.1016/j.asw.2020.100510

Barrot, J., & Gabinete, M. K. (2019). Complexity, accuracy, and fluency in the argumentative writing of ESL and EFL learners. *International Review of Applied Linguistics in Language Teaching*. https://doi.org/10.1515/iral-2017-0012

Beers, S. F., & Nagy, W. E. (2009). Syntactic complexity as a predictor of adolescent writing quality: Which measures? Which genre?. *Reading and Writing*, *22*(2), 185–200. https://doi.org/10.1007/s11145-007-9107-5

Biber, D., Gray, B., & Poonpon, K. (2011). Should we use characteristics of conversation to measure grammatical complexity in L2 writing development?. *TESOL Quarterly*, *45*(1), 5–35. https://doi.org/10.5054/tq.2011.244483

Bialystok, E. (1994). Analysis and control in the development of second language proficiency. *Studies in Second Language Acquisition*, *16*(2), 157–168. https://doi.org/10.1017/S0272263100012857

Blanchard, D., Tetreault, J., Higgins, D., Cahill, A., & Chodorow, M. (2013). TOEFL11: A corpus of non-native English. *ETS Research Report Series*, *2013*(2), i–15. https://doi.org/10.1002/j.2333-8504.2013.tb02331.x

Bulté, B., & Housen, A. (2012). Defining and operationalising L2 complexity. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (pp. 23–46). John Benjamins Publishing Company.

Bulté, B., & Housen, A. (2014). Conceptualizing and measuring short-term changes in L2 writing complexity. *Journal of Second Language Writing*, *26*, 42–65. https://doi.org/10.1016/j.jslw.2014.09.005

Canale, M., & Swain, M. (l980). Theoretical bases of communicative approaches to Second language teaching and testing. *Applied Linguistics*, *1*(1), 1–48.

Cao, Y., & Xiao, R. (2013). A multi-dimensional contrastive study of English abstracts by native and non-native writers. *Corpora*, *8*(2), 209–234. https://doi.org/10.3366/cor.2013.0041

Casal, J. E., & Kessler, M. (2020). Form and rhetorical function of phrase-frames in promotional writing: A corpus- and genre-based analysis. *System, 95.* https://doi.org/10.1016/j.system.2020.102370

Casal, J. E., & Lee, J. J. (2019). Syntactic complexity and writing quality in assessed first-year L2 writing. *Journal of Second Language Writing*, *44*, 51–62. https://doi.org/10.1016/j.jslw.2019.03.005

Chan, A. Y. (2010). Toward a taxonomy of written errors: Investigation into the written errors of Hong Kong Cantonese ESL learners. *TESOL Quarterly*, *44*(2), 295–319. https://doi.org/10.5054/tq.2010.219941

Chen, H., Xu, J., & He, B. (2014). Automated essay scoring by capturing relative writing quality. *The Computer Journal*, *57*(9), 1318–1330. https://doi.org/10.1093/comjnl/bxt117

Chuang, F.-Y., & Nesi, H. (2006). An analysis of formal errors in a corpus of L2 English produced by Chinese students. *Corpora*, *1*(2), 251–271. https://doi.org/10.3366/cor.2006.1.2.251

Coniam, D. (1999). An investigation into the use of word frequency lists in computing vocabulary profiles. *Hong Kong Journal of Applied Linguistics*, *4*(1), 103–123.

Crossley, S. A., & McNamara, D. S. (2009). Computational assessment of lexical differences in L1 and L2 writing. *Journal of Second Language Writing*, *18*(2), 119–135. https://doi.org/10.1016/j.jslw.2009.02.002

Crossley, S. A., & McNamara, D. (2012a). Detecting the first language of second language writers using automated indices of cohesion, lexical sophistication, syntactic complexity and conceptual knowledge. In S. Jarvis, & S. A. Crossley (Eds.), *Approaching language transfer through text classification: Axplorations in the detection-based approach* (pp. 106–126). Multilingual Matters.

Crossley, S. A., & McNamara, D. S. (2012b). Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication. *Journal of Research in Reading*, *35*(2), 115–135. https://doi.org/10.1111/j.1467-9817.2010.01449.x

Crossley, S. A., & McNamara, D. S. (2014). Does writing development equal writing quality? A computational investigation of syntactic complexity in L2 learners. *Journal of Second Language Writing*, *26*, 66–79. https://doi.org/10.1016/j.jslw.2014.09.006

Crossley, S. A., Salsbury, T., McNamara, D. S., & Jarvis, S. (2011). Predicting lexical proficiency in language learner texts using computational indices. *Language Testing*, *28*(4), 561–580. https://doi.org/10.1177/0265532210378031

Davies, Mark. (2004) *British National Corpus* (from Oxford University Press). Available online at https://www.english-corpora.org/bnc/.

Davies, Mark. (2008) *The Corpus of Contemporary American English (COCA)*. Available online at https://www.english-corpora.org/coca/.

De Angelis, G., & Jessner, U. (2012). Writing across languages in a bilingual context: A dynamic systems theory approach. In R. M. Manchón (Ed.), *L2 writing development: Multiple perspectives* (pp. 47–68). De Gruyter Mouton.

De Graaff, R., & Housen, A. (2009). Investigating the effects and effectiveness of L2 instruction. In M. H. Long, & C. J. Doughty (Eds.), *The handbook of language teaching* (pp. 726–755). Wiley-Blackwell.

Derwing, T. M., & Rossiter, M. J. (2003). The effects of pronunciation instruction on the accuracy, fluency, and complexity of L2 accented speech. *Applied language learning*, *13*(1), 1–17.

Eckstein, G., & Chang, R. H. (2022). How does the language control of L1 and L2 writers develop over time in first-year composition?. *Written Communication*, *39*(4), 600–629. https://doi.org/10.1177/07410883221099474

Eckstein, G., & Ferris, D. (2018). Comparing L1 and L2 texts and writers in first-year composition. *TESOL Quarterly*, *52*(1), 137–162. https://doi.org/10.1002/tesq.376

Educational Testing Service. (2020). Reliability and comparability of TOEFL iBT scores. *TOEFL iBT Research Insight*, *3*, 1–16.

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*(1), 59–84. https://doi.org/10.1017/S0272263104026130

Engber, C. A. (1995). The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second Language Writing*, *4*(2), 139–155. https://doi.org/10.1016/1060-3743(95)90004-7

Foster, P., & Wigglesworth, G. (2016). Capturing accuracy in second language performance: The case for a weighted clause ratio. *Annual Review of Applied Linguistics*, *36*, 98–116. https://doi.org/10.1017/S0267190515000082

Friginal, E., Li, M., & Weigle, S. C. (2014). Revisiting multiple profiles of learner compositions: A comparison of highly rated NS and NNS essays. *Journal of Second Language Writing*, *23*, 1–16. https://doi.org/10.1016/j.jslw.2013.10.001

Fritz, E., & Ruegg, R. (2013). Rater sensitivity to lexical accuracy, sophistication and range when assessing writing. *Assessing Writing*, *18*(2), 173–181. https://doi.org/10.1016/j.asw.2013.02.001

González, M. C. (2017). The contribution of lexical diversity to college-level writing. *TESOL Journal*, *8*(4), 899–919. https://doi.org/10.1002/tesj.342

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, *36*(2), 193–202. https://doi.org/10.3758/BF03195564

Granger, S. (1996). From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora. In K. Aijmer, B. Altenberg, & M. Johansson (Eds.), *Languages in contrast: Paper from a symposium on text-based cross-linguistic studies* (pp. 37–51). Lund University Press.

Granger, S., Dagneaux, E., Meunier, F., & Paquot, M. (Eds.). (2009). *International corpus of learner English Version 2*. Presses universitaires de Louvain.

Grant, L., & Ginther, A. (2000). Using computer-tagged linguistic features to describe L2 writing differences. *Journal of Second Language Writing*, *9*(2), 123–145. https://doi.org/10.1016/S1060-3743(00)00019-9

Gyllstad, H., Granfeldt, J., Bernardini, P., & Källkvist, M. (2014). Linguistic correlates to communicative proficiency levels of the CEFR: The case of syntactic complexity in written L2 English, L3 French and L4 Italian. *EUROSLA Yearbook*, *14*(1), 1–30. https://doi.org/10.1075/eurosla.14.01gyl

He, L., & Shi, L. (2012). Topical knowledge and ESL writing. *Language Testing*, *29*(3), 443–464. https://doi.org/10.1177/0265532212436659

Hinkel, E. (2009). The effects of essay topics on modal verb uses in L1 and L2 academic writing. *Journal of Pragmatics*, *41*(4), 667–683. https://doi.org/10.1016/j.pragma.2008.09.029

Housen, A., & Kuiken, F. (2009). Complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics*, *30*(4), 461–473. https://doi.org/10.1093/applin/amp048

Housen, A., Kuiken, F., & Vedder, I. (2012). *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA*. John Benjamins Publishing Company.

Hunt, K. W. (1965). Grammatical structures written at three grade levels. *National Council of Teachers of English Research Report No. 3*.

Jarvis, S. (2002). Short texts, best-fitting curves and new measures of lexical diversity. *Language Testing*, *19*(1), 57–84. https://doi.org/10.1191/0265532202lt220oa

Jarvis, S., & Crossley, S. A. (2012). *Approaching language transfer through text classification: Explorations in the detection-based approach*. Multilingual Matters.

Jarvis, S., Grant, L., Bikowski, D., & Ferris, D. (2003). Exploring multiple profiles of highly rated learner compositions. *Journal of Second Language Writing*, *12*(4), 377–403. https://doi.org/10.1016/j.jslw.2003.09.001

Jiang, J., Bi, P., & Liu, H. (2019). Syntactic complexity development in the writings of EFL learners: Insights from a dependency syntactically-annotated corpus. *Journal of Second Language Writing*, *46*. https://doi.org/10.1016/j.jslw.2019.100666

Johnson, M. D. (2017). Cognitive task complexity and L2 written syntactic complexity, accuracy, lexical complexity, and fluency: A research synthesis and meta-analysis. *Journal of Second Language Writing*, *37*, 13–38. https://doi.org/10.1016/j.jslw.2017.06.001

Johnson, M. D., Mercado, L., & Acevedo, A. (2012). The effect of planning sub-processes on L2 writing fluency, grammatical complexity, and lexical complexity. *Journal of Second Language Writing*, *21*(3), 264–282. https://doi.org/10.1016/j.jslw.2012.05.011

Kachru, B. B. (1985). Standard, codification and sociolinguistic realism: The English language in the outer circle. In R. Quirk, & H. Widdowson (Eds.), *English in the world: Teaching and learning the language and literatures* (pp. 11–30). Cambridge University Press.

Kessler, M., Ma, W., & Solheim, I. (2022). The effects of topic familiarity on text quality, complexity, accuracy, and fluency: A conceptual replication. *TESOL Quarterly*, *56*(4), 1163–1190. https://doi.org/10.1002/tesq.3096

Khushik, G. A., & Huhta, A. (2020). Investigating syntactic complexity in EFL learners' writing across common European framework of reference levels A1, A2, and B1. *Applied Linguistics*, *41*(4), 506–532. https://doi.org/10.1093/applin/amy064

Kim, J. Y. (2014). Predicting L2 writing proficiency using linguistic complexity measures: A corpus-based study. *English Teaching*, *69*(4), 27–51. https://doi.org/10.15858/engtea.69.4.201412.27

Kim, M., & Crossley, S. A. (2018). Modeling second language writing quality: A structural equation investigation of lexical, syntactic, and cohesive features in source-based and independent writing. *Assessing Writing*, *37*, 39–56. https://doi.org/10.1016/j.asw.2018.03.002

Kim, S., & Kessler, M. (2022). Examining L2 English university students' uses of lexical bundles and their relationship to writing quality. *Assessing Writing*, *51*. https://doi.org/10.1016/j.asw.2021.100589

Koizumi, R., & In'nami, Y. (2012). Effects of text length on lexical diversity measures: Using short texts with less than 200 tokens. *System*, *40*(4), 554–564. https://doi.org/10.1016/j.system.2012.10.012

Kuiken, F., & Vedder, I. (2007). Task complexity and measures of linguistic performance in L2 writing. *IRAL*, *45*, 261–284. https://doi.org/10.1515/iral.2007.012

Kuiken, F., & Vedder, I. (2019). Syntactic complexity across proficiency and languages: L2 and L1 writing in Dutch, Italian and Spanish. *International Journal of Applied Linguistics*, *29*(2), 192–210. https://doi.org/10.1111/ijal.12256

Kyle, K. (2016). *Measuring syntactic development in L2 writing: Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication* [Doctoral dissertation, Georgia State University]. ScholarWorks.

Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly*, *49*(4), 757–786. https://doi.org/10.1002/tesq.194

Kyle, K., & Crossley, S. (2017). Assessing syntactic sophistication in L2 writing: A usage-based approach. *Language Testing*, *34*(4), 513–535. https://doi.org/10.1177/0265532217712554

Kyle, K., & Crossley, S. A. (2018). Measuring syntactic complexity in L2 writing using fine-grained clausal and phrasal indices. *The Modern Language Journal*, *102*(2), 333–349. https://doi.org/10.1111/modl.12468

Kyle, K., Crossley, S., & Berger, C. (2018). The tool for the automatic analysis of lexical sophistication (TAALES): Version 2.0. *Behavior Research Methods*, *50*(3), 1030–1046. https://doi.org/10.3758/s13428-017-0924-4

Kyle, K., Crossley, S. A., & Jarvis, S. (2021). Assessing the validity of lexical diversity indices using direct judgements. *Language Assessment Quarterly*, *18*(2), 154–170. https://doi.org/10.1080/15434303.2020.1844205

Lan, G., Lucas, K., & Sun, Y. (2019). Does L2 writing proficiency influence noun phrase complexity? A case analysis of argumentative essays written by Chinese students in a first-year composition course. *System*, *85*. https://doi.org/10.1016/j.system.2019.102116

Larsen-Freeman, D. (1978). An ESL index of development. *TESOL Quarterly*, *12*(4), 439–448. https://doi.org/10.2307/3586142

Larsen-Freeman, D. (2009). Adjusting expectations: The study of complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics*, *30*(4), 579–589. https://doi.org/10.1093/applin/amp043

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, *16*(3), 307–322. https://doi.org/10.1093/applin/16.3.307

Lemmouh, Z. (2008). The relationship between grades and the lexical richness of student essays. *Nordic Journal of English Studies*, *7*(3), 163–180. https://doi.org/10.35360/njes.106

Levshina, N. (2015). *How to do linguistics with R*: *Data exploration and statistical analysis.* John Benjamins Publishing Company.

Li, H. (2015). Relationship between measures of syntactic complexity and judgments of EFL writing quality. In B. Ma, L. Cheng, H. He, L. Hale, & J. Zhang (Eds.), *Proceedings of 2015 youth academic forum on linguistics, literature, translation and culture* (pp. 216–222). The American Scholars Press.

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

Liu, J. (2008). L1 use in L2 vocabulary learning: Facilitator or barrier. *International Education Studies*, *1*(2), 65–69.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, *15*(4), 474–496. https://doi.org/10.1075/ijcl.15.4.02lu

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. https://doi.org/10.5054/tq.2011.240859

Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal*, *96*(2), 190–208. https://doi.org/10.1111/j.1540-4781.2011.01232_1.x

Lu, X. (2017). Automated measurement of syntactic complexity in corpus-based L2 writing research and implications for writing assessment. *Language Testing*, *34*(4), 493–511. https://doi.org/10.1177/0265532217710675

Lu, X., & Ai, H. (2015). Syntactic complexity in college-level English writing: Differences among writers with diverse L1 backgrounds. *Journal of Second Language Writing*, *29*, 16–27. https://doi.org/10.1016/j.jslw.2015.06.003

MacArthur, C. A., Jennings, A., & Philippakos, Z. A. (2019). Which linguistic features predict quality of argumentative writing for college basic writers, and how do those features change with instruction?. *Reading and Writing*, *32*(6), 1553–1574. https://doi.org/10.1007/s11145-018-9853-6

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk*. Lawrence Erlbaum Associates.

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

Mancilla, R. L., Polat, N., & Akcay, A. O. (2017). An investigation of native and nonnative English speakers' levels of written syntactic complexity in asynchronous online discussions. *Applied Linguistics*, *38*(1), 112–134. https://doi.org/10.1093/applin/amv012

Mansourizadeh, K., & Ahmad, U. K. (2011). Citation practices among non-native expert and novice scientific writers. *Journal of English for Academic Purposes*, *10*(3), 152–161. https://doi.org/10.1016/j.jeap.2011.03.004

Martínez, A. C. L. (2018). Analysis of syntactic complexity in secondary education EFL writers at different proficiency levels. *Assessing Writing*, *35*, 1–11. https://doi.org/10.1016/j.asw.2017.11.002

Mazgutova, D., & Kormos, J. (2015). Syntactic and lexical development in an intensive English for Academic Purposes programme. *Journal of second language writing*, *29*, 3–15. https://doi.org/10.1016/j.jslw.2015.06.004

McCarthy, P. M. (2005). *An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD)* [Doctoral dissertation, University of Memphis]. ProQuest.

McCarthy, P. M., & Jarvis, S. (2007). Vocd: A theoretical and empirical evaluation. *Language Testing*, *24*(4), 459–488. https://doi.org/10.1177/0265532207080767

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, *42*(2), 381–392. https://doi.org/10.3758/BRM.42.2.381

McKee, G., Malvern, D., & Richards, B. (2000). Measuring vocabulary diversity using dedicated software. *Literary and Linguistic Computing*, *15*(3), 323–338. https://doi.org/10.1093/llc/15.3.323

McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2010). Linguistic features of writing quality. *Written Communication*, *27*(1), 57–86. https://doi.org/10.1177/0741088309351547

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press.

Midway, S., Robertson, M., Flinn, S., & Kaller, M. (2020). Comparing multiple comparisons: Practical guidance for choosing the best multiple comparisons test. *PeerJ*, *8*.

Mostafa, T., & Crossley, S. A. (2020). Verb argument construction complexity indices and L2 writing quality: Effects of writing tasks and prompts. *Journal of Second Language Writing*, *49*. https://doi.org/10.1016/j.jslw.2020.100730

Murakami, A., Granger, S., Gaëtanelle, S., & Meunier, F. (2013). Cross-linguistic influence on the accuracy order of L2 English grammatical morphemes. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *Twenty years of learner corpus research: Looking back, moving ahead* (pp. 325–334). Presses universitaires de Louvain.

Nakamaru, S. (2010). Lexical issues in writing center tutorials with international and US-educated multilingual writers. *Journal of Second Language Writing*, *19*(2), 95–113. https://doi.org/10.1016/j.jslw.2010.01.001

Norris, J. M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics*, *30*(4), 555–578. https://doi.org/10.1093/applin/amp044

Ortega, L. (1995). The effect of planning in L2 Spanish oral narratives. *Studies in Second Language Acquisition*, *21*, 108–148.

Ortega, L. (2003). Syntactic complexity measures and their relationship to l2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics*, *24*(4), 492–518. https://doi.org/10.1093/applin/24.4.492

Ortega, L. (2015). Syntactic complexity in L2 writing: Progress and expansion. *Journal of Second Language Writing*, *29*, 82–94. https://doi.org/10.1016/j.jslw.2015.06.008

Pallotti, G. (2009). CAF: Defining, refining and differentiating constructs. *Applied Linguistics*, *30*(4), 590–601. https://doi.org/10.1093/applin/amp045

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

Polio, C. (2017). Second language writing development: A research agenda. *Language Teaching*, *50*(2), 261–275. https://doi.org/10.1017/S0261444817000015

Polio, C., & Shea, M. C. (2014). An investigation into current measures of linguistic accuracy in second language writing research. *Journal of Second Language Writing*, *26*, 10–27. https://doi.org/10.1016/j.jslw.2014.09.003

Qin, W., & Uccelli, P. (2016). Same language, different functions: A cross-genre analysis of Chinese EFL learners' writing performance. *Journal of Second Language Writing*, *33*, 3–17. https://doi.org/10.1016/j.jslw.2016.06.001

R Core Team. (2022). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*. https://www.R-project.org/.

Rahayu, F. E. S., Utomo, A., & Setyowati, R. (2021). Syntactic and lexical complexity of undergraduate students' essays: a comparison study between L1 and L2 writings. *Indonesian Journal of English Language Teaching and Applied Linguistics*, *5*(2), 251–263.

Rankin, T. (2012). The transfer of V2: Inversion and negation in German and Dutch learners of English. *International Journal of Bilingualism*, *16*(1), 139–158. https://doi.org/10.1177/1367006911405578

Read, J. (2000). *Assessing vocabulary*. Cambridge University Press.

Richards, B. (1987). Type/token ratios: What do they really tell us?. *Journal of Child Language*, *14*(2), 201–209. https://doi.org/10.1017/S0305000900012885

Ringbom, H. (2007). *Cross-linguistic similarity in foreign language learning*. Multilingual Matters.

Ringbom, H., & Jarvis, S. (2009). The importance of cross-linguistic similarity in foreign language learning. In M. H. Long, & C. J. Doughty (Eds.), *The handbook of language teaching* (pp. 106–118). Wiley-Blackwell.

Robinson, P. (2001). Task complexity, task difficulty, and task production: Exploring interactions in a componential framework. *Applied Linguistics*, *22*(1), 27–57. https://doi.org/10.1093/applin/22.1.27

Romano, F. (2019). Grammatical accuracy in EAP writing. *Journal of English for Academic Purposes*, *41*. https://doi.org/10.1016/j.jeap.2019.100773

Salazar, D. (2014). *Lexical bundles in native and non-native scientific writing: Applying a corpus-based study to language teaching*. John Benjamins Publishing Company.

Sauder, D. C., & DeMars, C. E. (2019). An updated recommendation for multiple comparisons. *Advances in Methods and Practices in Psychological Science*, *2*(1), 26–44. https://doi.org/10.1177/2515245918808784

Singh, C. K. S., Singh, A. K. J., Razak, N. Q. A., & Ravinthar, T. (2017). Grammar errors made by ESL tertiary students in writing. *English Language Teaching*, *10*(5), 16–27. http://doi.org/10.5539/elt.v10n5p162013

Skehan, P. (1996). A framework for the implementation of task-based instruction. *Applied Linguistics*, *17*(1), 38–62. https://doi.org/10.1093/applin/17.1.38

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

Skehan, P. (2003). Task-based instruction. *Language Teaching*, *36*(1), 1–14. https://doi.org/10.1017/S026144480200188X

Skehan, P., & Foster, P. (1997). Task type and task processing conditions as influences on foreign language performance. *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. 183–205). Cambridge University Press.

Taguchi, N., Crawford, W., & Wetzel, D. Z. (2013). What linguistic features are indicative of writing quality? A case of argumentative essays in a college composition program. *TESOL Quarterly*, *47*(2), 420–430.

Tedick, D. J. (1990). ESL writing assessment: Subject-matter knowledge and its impact on performance. *English for Specific Purposes*, *9*(2), 123–143. https://doi.org/10.1016/0889-4906(90)90003-U

Tetreault, J., Blanchard, D., & Cahill, A. (2013). A report on the first native language identification shared task. In J. Tetreault, J. Burstein, & C. Leacock (Eds.), *Proceedings of the eighth workshop on innovative use of NLP for building educational applications* (pp. 48–57). Association of Computational Linguistics.

Tomczak, M., & Tomczak, E. (2014). The need to report effect size estimates revisited: An overview of some recommended measures of effect size. *Trends in Sport Sciences*, *21*(1).

Uysal, H. H. (2008). Tracing the culture behind writing: Rhetorical patterns and bidirectional transfer in L1 and L2 essays of Turkish writers in relation to educational context. *Journal of Second Language Writing*, *17*(3), 183–207. https://doi.org/10.1016/j.jslw.2007.11.003

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. 93–116). Cambridge University Press.

van Vuuren, S. (2013). Information structural transfer in advanced Dutch EFL writing: A cross-linguistic longitudinal study. *Linguistics in the Netherlands*, *30*(1), 173–187. https://doi.org/10.1075/avt.30.13van

van Weijen, D., van den Bergh, H., Rijlaarsdam, G., & Sanders, T. (2009). L1 use during L2 writing: An empirical study of a complex phenomenon. *Journal of Second Language Writing*, *18*(4), 235–250. https://doi.org/10.1016/j.jslw.2009.06.003

Verspoor, M., Schmid, M. S., & Xu, X. (2012). A dynamic usage-based perspective on L2 writing. *Journal of Second Language Writing*, *21*(3), 239–263. https://doi.org/10.1016/j.jslw.2012.03.007

Vo, P. D., & Barrot, J. S. (2022). Complexity, accuracy, and fluency in L2 writing across proficiency levels: A matter of L1 background?. *Assessing Writing*, *54*. https://doi.org/10.1016/j.asw.2022.100673

Vyatkina, N., Hirschmann, H., & Golcher, F. (2015). Syntactic modification at early stages of L2 German writing development: A longitudinal learner corpus study. *Journal of Second Language Writing*, *29*, 28–50. https://doi.org/10.1016/j.jslw.2015.06.006

Wee, R., Sim, J., & Jusoff, K. (2010). Verb-form errors in EAP writing. *Educational Research and Reviews*, *5*(1), 16–23. https://doi.org/10.5897/ERR.9000408

Wigglesworth, G., & Storch, N. (2009). Pair versus individual writing: Effects on fluency, complexity and accuracy. *Language Testing*, *26*(3), 445–466. https://doi.org/10.1177/0265532209104670

Wolfe-Quintero, K., Inagaki, S., & Kim, H.-Y. (1998). *Second language development in writing: Measures of fluency, accuracy, & complexity*. University of Hawaii Press.

Yang, W. (2014). *Mapping the relationships among the cognitive complexity of independent writing tasks, L2 writing quality, and complexity, accuracy and fluency of l2 writing* [Doctoral dissertation, Georgia State University]. ScholarWorks.

Yang, W., & Kim, Y. (2020). The effect of topic familiarity on the complexity, accuracy, and fluency of second language writing. *Applied Linguistics Review*, *11*(1), 79–108. https://doi.org/10.1515/applirev-2017-0017

Yang, W., Lu, X., & Weigle, S. C. (2015). Different topics, different discourse: Relationships among writing topic, measures of syntactic complexity, and judgments of writing quality. *Journal of Second Language Writing*, *28*, 53–67. https://doi.org/10.1016/j.jslw.2015.02.002

Yigit, S., & Mendes, M. (2018). Which effect size measure is appropriate for one-way and two-way ANOVA models? A Monte Carlo simulation study. *Revstat-Statistical Journal*, *16*(3), 295–313.

Yoon, H. J. (2017). Linguistic complexity in L2 writing revisited: Issues of topic, proficiency, and construct multidimensionality. *System*, *66*, 130–141. https://doi.org/10.1016/j.system.2017.03.007

Yoon, H. J., & Polio, C. (2017). The linguistic development of students of English as a second language in two written genres. *TESOL Quarterly*, *51*(2), 275–301. https://doi.org/10.1002/tesq.296

Yu, G. (2010). Lexical diversity in writing and speaking task performances. *Applied Linguistics*, *31*(2), 236–259. https://doi.org/10.1093/applin/amp024

Yuan, F., & Ellis, R. (2003). The effects of pre-task planning and on-line planning on fluency, complexity and accuracy in L2 monologic oral production. *Applied Linguistics*, *24*(1), 1–27. https://doi.org/10.1093/applin/24.1.1

Zenker, F., & Kyle, K. (2021). Investigating minimum text lengths for lexical diversity indices. *Assessing Writing*, *47*. https://doi.org/10.1016/j.asw.2020.100505

Zhang, X., & Lu, X. (2022). Revisiting the predictive power of traditional vs. fine-grained syntactic complexity indices for L2 writing quality: The case of two genres. *Assessing Writing*, *51*. https://doi.org/10.1016/j.asw.2021.100597

APPENDIX A: CODING GUIDELINES

The following coding guidelines are adapted from adapted from Yoon and Polio (2017). Read the essay sentence by sentence and highlight errors (or the phrases containing the errors) based on their error types. If a phrase contains 2 different error types, highlight the errors using different colors. In case an error doesn't belong to one of the error types below, make a comment to note it down. Formatting issues should be ignored.

1. Incomprehensible (red)

- Incomprehensible sentences or clauses with unclear intended meaning (only when they are grammatically problematic)
    - *Do not make undermeasiring in the communication because to can use internet*.

2. Syntactic errors (orange)

- Incorrect word order
    - *I didn't know what should I [I should] do*.
- Sentence fragments
    - *When we were tired*.
- Run-on sentences and comma splices
    - *I dance, I sang*.
- Missing constituents
    - *Studied hard. I like*.
    - *I put the book.*
- Extra verbs or subjects in a clause

- *He walks listens to music*.

- Infelicitous uses of relative clauses
    - *I like the book who [that] I read*.

3. Morphological errors (yellow)

- Incorrect uses of word form (including POS)
    - *He choice [chose] a very good restaurant*.
- Subject-verb agreement
    - *She bring [brings] her homework.*
- Plurals
    - *I have many story [stories] about her*.
- Genitive
    - *My friends [friend's] mom is very nice*.
- Articles
    - *She wants to buy new car*.
    - *She has a three children*.
- Double negatives
    - *There is not no one to take it*.
- Wrong pronouns in terms of gender or case
    - *James said her [his] mom is nice. I like she [her]*.
- Verb form problems including tense-aspect, passive voice, missing or extra *to*-infinitives, modals
    - *They teaching today*.
    - *The event was happened*.

- *I can <u>to</u> travel during the break.*

4. Preposition errors (blue)

- All infelicitous uses of prepositions—missing, extra, or wrong prepositions
    - *I came the U.S. I made friends <u>at</u> there.*
    - *Culture us the way <u>in</u> [of] living.*

5. Spelling errors (green)

- All misspellings according to standard British/American English orthography
    - *Recently, some <u>psycologyst</u> proclaimed that <u>prefering</u> the old things make people can't get success.*

APPENDIX B: BETWEEN-L1 DIFFERENCES IN SYNTACTIC COMPLEXITY

Table 19

Between-L1 Differences in Syntactic Complexity in Medium Score Level

	Arabic	Chinese	French	German	Hindi	Italian	Japanese	Korean	Spanish	Telugu	Turkish
Arabic	-	MLS (p = .01) CP/T (p = .006) CP/C (p = .02)	CP/T (p = .02) CP/C (p = .005)	T/S (p = .001) C/S (p = .002) MLT (p = .02) CN/T (p = .009) VP/T (p = .01)		CP/T (p = .001) CP/C (p < .001)	MLS (p < .001) T/S (p < .04) C/S (p = .001) MLT (p < .001) C/T (p = .02) DC/T (p = .02) CP/T (p = .001) CN/T (p < .001) VP/T (p = .001) MLC (p < .001) CN/C (p = .03)	MLS (p < .001) T/S (p < .001) C/S (p < .001) MLT (p < .001) C/T (p = .002) CT/T (p = .04) DC/T (p = .001) CP/T (p = .01) CN/T (p = .001) VP/T (p < .001) DC/C (p = .009)		MLS (p = .04) MLT (p = .001) C/T (p = .009) CT/T (p = .04) DC/T (p = .004) CN/T (p = .001) VP/T (p = .02)	

Table 19 (Continued)

	Arabic	Chinese	French	German	Hindi	Italian	Japanese	Korean	Spanish	Telugu	Turkish
Chinese		-			MLS (p = .001) C/T (p < .001) CT/T (p = .007) DC/T (p < .001) CP/T (p < .001) CN/T (p < .001) VP/T (p = .001) DC/C (p = .004)	T/S (p = .03) CT/T (p = .04)	MLS (p = .01) MLT (p = .01)	MLS (p = .001) T/S (p = .045) C/S (p = .001) MLT (p = .01) C/T (p = .004) DC/T (p = .03) VP/T (p = .001)	CP/T (p = .049)	MLS (p < .001) C/S (p < .001) MLT (p < .001) C/T (p < .001) CT/T (p < .001) DC/T (p < .001) CP/T (p < .001) CN/T (p < .001) VP/T (p < .001) MLC (p = .004) DC/C (p < .001) CN/C (p < .001)	MLS (p = .04) C/S (p = .03) CP/T (p = .04)

Table 19 (Continued)

	Arabic	Chinese	French	German	Hindi	Italian	Japanese	Korean	Spanish	Telugu	Turkish
French			-		MLS (p = .01) C/T (p = .04) CP/T (p = .001) CN/T (p = .002) VP/T (p = .04) CP/C (p = .02) CN/C (p = .007)		MLT (p = .002) VP/T (p = .01)	MLS (p = .003) C/S (p = .001) MLT (p = .002) C/T (p = .001) DC/T (p = .001) VP/T (p = .001) DC/C (p < .001)		MLS (p < .001) C/S (p < .001) MLT (p < .001) C/T p < .001) CT/T (p = .001) DC/T (p < .001) CP/T (p < .001) CN/T (p < .001) VP/T (p < .001) MLC (p = .001) DC/C (p = .03) CP/C (p = .04) CN/C (p < .001)	

Table 19 (Continued)

	Arabic	Chinese	French	German	Hindi	Italian	Japanese	Korean	Spanish	Telugu	Turkish
German				-	C/S ($p < .001$) C/T ($p < .001$) DC/T ($p < .001$) CP/T ($p = .005$) CN/T ($p < .001$) VP/T ($p < .001$) CN/C ($p = .045$)	T/S ($p < .001$) C/S ($p < .001$)	MLT ($p = .04$)	MLT ($p < .048$)	T/S ($p = .001$) C/S ($p = .001$) CN/T ($p = .048$) VP/T ($p = .02$)	T/S ($p = .02$) C/S ($p < .001$) MLT ($p < .001$) C/T ($p < .001$) CT/T ($p = .01$) DC/T ($p < .001$) CP/T ($p < .001$) CN/T ($p < .001$) VP/T ($p < .001$) DC/C ($p = .001$) CN/C ($p < .001$)	T/S ($p = .006$) C/S ($p < .001$) C/T ($p < .04$) CN/T ($p = .048$) VP/T ($p = .04$)

Table 19 (Continued)

	Arabic	Chinese	French	German	Hindi	Italian	Japanese	Korean	Spanish	Telugu	Turkish
Hindi					-	C/T (p = .01) DC/T (p = .02) CP/T (p < .001) CN/T (p = .003) VP/T (p = .002) CP/C (p < .001)	MLS (p < .001) C/S (p < .001) MLT (p = .02) C/T (p < .001) CT/T (p = .01) DC/T (p < .001) CP/T (p < .001) CN/T (p < .001) VP/T (p < .001) DC/C (p < .001) CN/C (p = .001)	MLS (p < .001) C/S (p < .001) MLT (p = .02) C/T (p < .001) CT/T (p < .001) DC/T (p < .001) CP/T (p < .001) CN/T (p < .001) VP/T (p < .001) DC/C (p < .001)	CN/T (p = .02)		

Table 19 (Continued)

	Arabic	Chinese	French	German	Hindi	Italian	Japanese	Korean	Spanish	Telugu	Turkish
Italian						-	MLS ($p < .001$) T/S ($p < .001$) C/S ($p < .001$) MLT ($p < .001$) C/T ($p = .005$) DC/T ($p = .02$) CN/T ($p < .001$) VP/T ($p = .006$) MLC ($p = .04$)	MLS ($p < .001$) T/S ($p < .001$) C/S ($p < .001$) MLT ($p < .001$) C/T ($p < .001$) CT/T ($p < .001$) DC/T ($p < .001$) CN/T ($p < .001$) VP/T ($p < .001$) DC/C ($p < .001$) CP/C ($p = .02$)	CP/T ($p = .003$) CP/C ($p = .004$)	MLS ($p < .001$) C/S ($p < .001$) MLT ($p < .001$) C/T ($p < .001$) DC/T ($p < .001$) CP/T ($p < .001$) CN/T ($p < .001$) VP/T ($p < .001$) MLC ($p = .02$) DC/C ($p = .009$) CP/C ($p < .001$) CN/C ($p < .001$)	CP/T ($p = .006$) CP/C ($p = .01$)

Table 19 (Continued)

	Arabic	Chinese	French	German	Hindi	Italian	Japanese	Korean	Spanish	Telugu	Turkish
Japanese							-		MLS ($p < .03$) C/S ($p < .001$) MLT ($p < .001$) C/T ($p = .001$) DC/T ($p = .001$) CP/T ($p = .008$) CN/T ($p < .001$) VP/T ($p < .001$) MLC ($p = .047$) DC/C ($p = .04$)	MLS ($p < .001$) C/S ($p < .001$) MLT ($p < .001$) C/T ($p < .001$) CT/T ($p < .001$) DC/T ($p < .001$) CP/T ($p < .001$) CN/T ($p < .001$) VP/T ($p < .001$) MLC ($p < .001$) DC/C ($p < .001$) CN/C ($p < .001$)	MLS ($p < .001$) C/S ($p < .001$) C/T ($p = .007$) DC/T ($p = .009$) CP/T ($p = .01$) CN/T ($p = .005$) VP/T ($p = .007$)

Table 19 (Continued)

	Arabic	Chinese	French	German	Hindi	Italian	Japanese	Korean	Spanish	Telugu	Turkish
Korean								-	MLS ($p < .02$) T/S ($p = .001$) C/S ($p < .001$) MLT ($p < .001$) C/T ($p < .001$) CT/T ($p = .007$) DC/T ($p < .001$) CN/T ($p < .001$) VP/T ($p < .001$) DC/C ($p < .001$)	MLS ($p < .001$) C/S ($p < .001$) MLT ($p < .001$) C/T ($p < .001$) CT/T ($p < .001$) DC/T ($p < .001$) CP/T ($p < .001$) CN/T ($p < .001$) VP/T ($p < .001$) MLC ($p = .001$) DC/C ($p < .001$) CN/C ($p = .001$)	MLS ($p < .001$) T/S ($p < .003$) C/S ($p < .001$) C/T ($p = .001$) DC/T ($p = .001$) CN/T ($p = .009$) VP/T ($p = .001$) DC/C ($p = .01$)

Table 19 (Continued)

	Arabic	Chinese	French	German	Hindi	Italian	Japanese	Korean	Spanish	Telugu	Turkish
Spanish									-	C/S (p = .001) MLT (p < .001) C/T (p < .001) DC/T (p < .001) CP/T (p = .005) CN/T (p < .001) VP/T (p < .001) CN/C (p = .004)	

Table 19 (Continued)

	Arabic	Chinese	French	German	Hindi	Italian	Japanese	Korean	Spanish	Telugu	Turkish
Telugu										-	CT/T ($p < .001$) DC/T ($p = .04$) CN/T ($p = .008$) MLC ($p = .02$) DC/C ($p = .002$) CN/C ($p < .001$)
Turkish											-

Table 20

Between-L1 Differences in Syntactic Complexity in High Score Level

	Arabic	Chinese	French	German	Hindi	Italian	Japanese	Korean	Spanish	Telugu	Turkish
Arabic	-	DC/C (p = .02)	CP/T (p = .02)	DC/C (p = .01)			DC/C (p = .03)	DC/C (p = .005)		CN/C (p = .04)	DC/C (p = .001)
Chinese		-			C/T (p = .03) DC/T (p = .01) CP/T (p = .002) CN/T (p = .01)	C/T (p = .046) DC/T (p = .04)			C/T (p < .001) DC/T (p < .001) VP/T (p = .008) DC/C (p = .02)	C/T (p = .049) DC/T (p = .03) CN/T (p = .02)	
French			-		CP/T (p = .001) CN/T (p = .02) MLC (p = .03) CP/C (p = .002) CN/C (p = .001)			MLS (p = .02) C/S (p < .001) C/T (p = .03)	MLS (p = .02) C/S (p = .02)	MLS (p = .048) MLT (p = .04) CN/T (p = .03) MLC (p = .004) CN/C (p < .001)	
German				-	MLS (p = .01) C/S (p = .04) MLT (p = .03) CP/T (p = .002) CN/T (p = .01) DC/C (p = .049) CN/C (p = .001)	MLS (p = .003) C/S (p = .006) MLT (p = .048) CN/T (p = .008)			MLS (p < .001) C/S (p < .001) MLT (p = .006) C/T (p < .001) DC/T (p < .001) VP/T (p < .001) DC/C (p = .005)	MLS (p = .02) MLT (p = .02) CN/T (p = .02) CN/C (p < .001)	

Table 20 (Continued)

	Arabic	Chinese	French	German	Hindi	Italian	Japanese	Korean	Spanish	Telugu	Turkish
Hindi					-			MLS (p = .001) C/S (p = .002) MLT (p = .005) C/T (p = .01) DC/T (p = .01) CP/T (p = .02) CN/T (p = .009) VP/T (p = .01) DC/C (p = .03)	CP/T (p = .04) CP/C (p = .04) CN/C (p = .007)		CP/T (p = .01) DC/C (p = .002)
Italian						-		MLS (p < .001) T/S (p = .04) C/S (p < .001) MLT (p = .001) C/T (p = .005) DC/T (p = .02) CN/T (p = .04) VP/T (p = .002)		CN/C (p = .04)	
Japanese							-		C/S (p = .005) DC/C (p = .03)	MLS (p = .04) MLT (p = .045)	

Table 20 (Continued)

	Arabic	Chinese	French	German	Hindi	Italian	Japanese	Korean	Spanish	Telugu	Turkish
Korean								-	MLS (p < .001) T/S (p = .007) C/S (p < .001) MLT (p < .001) C/T (p < .001) DC/T (p < .001) VP/T (p < .001) DC/C (p = .004)	MLS (p = .002) C/S (p = .01) MLT (p = .004) C/T (p = .03) DC/T (p = .03) CN/T (p = .02) VP/T (p = .03) DC/C (p = .04)	
Spanish									-	MLC (p = .02) CN/C (p < .001)	DC/C (p < .001)
Telugu										-	DC/C (p = .006)
Turkish											-