SYNTACTIC COMPLEXITY AS A PREDICTOR OF SECOND LANGUAGE WRITING PROFICIENCY AND WRITING QUALITY

By

Ji-Hyun Park

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Second Language Studies—Doctor of Philosophy

2017

ABSTRACT

SYNTACTIC COMPLEXITY AS A PREDICTOR OF SECOND LANGUAGE WRITING PROFICIENCY AND WRITING QUALITY

By

Ji-Hyun Park

Syntactic (i.e., grammatical) complexity refers to the *range* and the degree of *sophistication* of the forms that appear in language production (Ortega, 2003). This concept has long been regarded as an important construct of language proficiency and has been actively investigated in the field of second language (L2) writing. Syntactic complexity is multidimensional in nature, and there are a variety of measures that tap into different dimensions of the construct. Widely used measures of complexity (e.g., mean length of T-unit and the number of clauses per T-unit) capture a relative degree of sophistication, but do not provide measures of participants' command of a range of diverse syntactic structures. In contrast, in L2 assessment, grammatical knowledge is often evaluated in terms of both syntactic elaboration and structural variety (Rimmer, 2006). To address the gap, the present study proposes new measures that tap into the diversity dimension of syntactic complexity: types and type/token frequency of verb-argument constructions (VACs). The present study investigates whether the proposed diversity measures of syntactic complexity, in combination with currently used measures of elaboration, accurately predict L2 written proficiency and writing quality. Specific research questions that guide the study are as follows: (1) Does the syntactic complexity of Korean EFL learners' writing production, as measured by various quantitative complexity measures, function as an indicator of proficiency? In addition, does adding diversity measures increase the predictive power of syntactic complexity in discriminating proficiency levels? (2) How do different syntactic

complexity measures relate to subjective ratings of writing quality judged by human raters? Which measure(s) best predict writing quality? (3) How do raters interpret the notion of syntactic complexity that appear on *Language Use* scale of a given analytic writing rubric?

Essays were collected from 390 Korean EFL learners and analyzed using corpus analytic tools. Fourteen elaboration measures were calculated using Syntactic Complexity Analyzer, an automated computational tool developed by Lu (2010). For the diversity measures, all instances of VACs in the participants' essays were retrieved and analyzed using a part-of-speech tagging tool and a concordance tool. Thirteen VAC patterns (e.g., verb + direct object, verb + indirect object + direct object, verb + direct object + object predicative, etc.) and their sub-patterns were identified based on findings in usage-based approaches to grammar, namely, construction grammar and corpus-based descriptive grammar. Then the distribution and the number of VAC types and type/token frequency of VACs were examined. Participants' proficiency levels were independently measured by a cloze test, and the quality of their essays was evaluated by human raters. The empirical results of the study indicated that measures of syntactic complexity functioned as predictors that discriminate among different proficiency levels, and adding diversity measures of complexity increased the predictive power. The diversity measures were also found to be strong predictors of human-rated writing quality, which lend support to the use of the diversity measures in this area of research. Qualitative data obtained from the rater interviews showed that notions of grammatical complexity as interpreted by raters generally overlap with the notion of syntactic complexity in SLA. However, variability was found in the interpretations between raters.

ACKNOWLEDGMENTS

First of all I would like to sincerely thank my advisor Dr. Charlene Polio for her guidance during my years in the SLS program. I am very thankful for her extraordinary support and encouragement throughout the dissertation project. My special thanks go to the committee members, Dr. Susan Gass, Dr. Patti Spinner, Dr. Paula Winke, and Dr. Aline Godfroid, for their time reading my dissertation and helpful comments. I also want to thank my advisors at Seoul National University, Professor Hyun-Kwon Yang and Hyunkee Ahn, for being great mentors through my first years of research. Their commitment and dedication to teaching and research has always been an inspiration to me.

This dissertation would not have been possible without the help of a number of people: the instructors in Korea who helped data collection, the students who participated in the study, and the raters who read the student essays. I also gratefully acknowledge the financial support from a Dissertation Completion Fellowship from the College of Arts and Letters and a Special College Research Abroad Money award from the Graduate School at Michigan State University.

I also need to thank my colleagues and friends who helped make my time in the doctoral program more bearable and enjoyable. Special thanks go to Ina Choi, Yaqiong Cui, Talip Gonulal, Lorena Valmori, Susie Kim, and Unhee Ju. I also thank Magda Tigchelaar for reading my dissertation.

Last but not least, I am grateful to my family for their love and support. I thank my parents and parents-in-law who have always believed in me and supported my decisions in every way. My most heartfelt gratitude goes to my husband, Jongwon Shin. Without his encouragement, care and unfailing love, this work would not have been completed.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

What it means to be a proficient language user and how to describe and measure learners' proficiency are two major questions that have been at the core of many studies in second language acquisition (SLA) and applied linguistics (Housen & Kuiken, 2009). There is now a shared belief among researchers and practitioners that second language (L2) proficiency, both oral and written, is a multidimensional rather than unitary construct. This multidimensionality has been captured by three constructs, namely, complexity, accuracy, and fluency (Ellis, 2003; Housen, Kuiken, & Vedder, 2012; Norris & Ortega, 2009; Skehan, 1998), which have become recognized as "principal and basic dimensions of L2 performance, proficiency, and development" (Bulté & Housen, 2014, p.13). Originating from L1 research and first introduced by Skehan (1998) in an L2 model, these three constructs have emerged as research variables in the field of SLA over the past 25 years (Housen & Kuiken, 2009).

Among the three constructs, complexity—especially syntactic complexity (also called grammatical complexity)—has a long history in the research on L2 writing development (Biber, Gray, & Poonpon, 2011). Often defined as "the range of forms that surface in language production and the degree of sophistication of such forms" (Ortega, 2003, p. 492), complexity has been recognized as an important construct in L2 writing teaching and research. Researchers have assumed that learner language becomes more complex as learners progress and have viewed increased complexity as an indication of language development or proficiency. Accordingly, establishing and scrutinizing measures of syntactic complexity has become common.

Developing objective methods to assess language proficiency has been one of the main goals in the L2 assessment field. Grammatical competence is one aspect of communicative

competence (Bachman, 1990; Canale & Swain, 1980) and is central to describing test-taker performance (Rimmer, 2006). In addition, in describing grammatical competence, grammatical complexity (complexity of form and structure) is considered to be crucial (Rimmer, 2006). For example, rubrics used to rate the speaking or writing performance of L2 learners (e.g., TOEFL writing rubrics, IELTS writing band descriptors) often illustrate the use of a variety of syntactic structures or sentence forms as a measure of test-takers' language use.

Although research in both SLA and L2 assessment pursue a similar goal, few attempts have been made to compare and contrast how the construct of syntactic (grammatical) complexity is interpreted and operationalized in each field. In the present study, I attempt to build a connection between the two fields so that findings and practices in these areas can inform each other. Specifically, this study aims to study how syntactic or grammatical complexity has been operationalized in each field, critically review the measures of complexity and examine the relationship between measures in the two fields, and propose new measures to fill the gap. In addition, I investigate whether the proposed measures, together with conventional complexity measures that have been used in SLA, can be indicative of L2 writing proficiency and writing quality as judged by human raters.

CHAPTER 1: BACKGROUND

In this chapter, I first examine how syntactic complexity has been defined in SLA and L2 writing research. This is followed by a summary of syntactic complexity measures that have been frequently used in SLA. I then introduce previous studies that employed these measures to identify how and for what purposes these measures have been used in SLA. I focus in particular on studies that investigated the relationship between syntactic complexity and L2 development and proficiency levels. The following section contains a review of studies on L2 assessment. I examine how grammatical development has been viewed and measured, and how the test-takers' performance is interpreted in relation to these measures. Then I review studies that investigate the link between the syntactic complexity measures used in the SLA field and writing performance assessed by human raters. I note that currently used measures do not capture the diversity dimension of complexity and that a mismatch exists between the interpretations of the construct in SLA and in L2 assessment. At the end of this chapter, I propose new measures of syntactic complexity, informed by findings in usage-based linguistics, in an attempt to fill the gap.

1.1 Complexity in SLA research

Research on complexity and complex systems has flourished since the 1990s in various disciplines such as the natural, social, and psychological sciences as well as language sciences (Bulté & Housen, 2014). Although no consensus has yet emerged on the definition of complexity, this construct is commonly understood across the disciplines as a property or entity in terms of "(1) the number and the nature of the discrete components that the entity consists of, and (2) the number and the nature of the relationship between the constituent components" (Bulté & Housen, 2012, p.22). For example, in the language sciences, including SLA and

applied linguistics, complexity is often defined in terms of the number and the nature of language components and the combinations thereof, as reflected in some traditional working definition of complexity such as "using a wide range of structures and vocabulary" (Lennon, 1990, p.390) or "[t]he extent to which the language produced in performing a task is elaborate and varied" (Ellis, 2003, p.340).

Dictionaries define complexity as "(1) the quality or state of not being simple: the quality or state of being complex; (2) a part of something that is complicated or hard to understand (Merriam-Webster Dictionary)." In the field of SLA, researchers have acknowledged these two meanings of complexity by distinguishing absolute and relative complexity (Pallotti, 2015). This distinction is also referred to as objective and subjective. Absolute or objective complexity refers to learner-independent linguistic properties, while relative or subjective complexity is a language-user or learner-dependent concept related to learners' cognitive abilities. Bulté and Housen used the term (cognitive) difficulty to refer to the latter concept (subjective or relative complexity) and reserved the term complexity for L2 linguistic complexity.

Many researchers have pointed out the difficulty in defining complexity in SLA studies (Housen & Kuiken, 2009; Pallotti, 2015; Vyatkina, Hirschmann & Golcher, 2015). Several reasons account for this. First, the term complexity has been used to refer to both features of a communicative task that learners perform (task complexity) and language produced by learners in the field (L2 complexity). L2 complexity can be, again, interpreted as either absolute (also called objective) complexity or relative complexity (subjective complexity, cognitive complexity, or difficulty), as described above. In addition, complexity can be observed in various language subsystems such as vocabulary, morphology and syntax, which makes it hard to treat as a single construct.

Researchers have often failed to capture the complex and multi-faceted nature of the construct when defining complexity, and used very general and vague terms in defining and operationalizing complexity (Bulté & Housen, 2012). Recently, in an attempt to advance the understanding of the construct, several researchers have tried to describe complexity from a more comprehensive and systematic perspective (e.g., Bulté & Housen 2012; Norris & Ortega, 2009; Ortega, 2012; Pallotti, 2009). One of the most recent attempts to conceptualize the notion of complexity in SLA is the work by Bulté and Housen (2012), who classified components of complexity at several levels. In their taxonomic model of L2 complexity, the authors first distinguished difficulty from complexity, and further categorized complexity. In a broad sense, (L2) complexity consists of linguistic, discourse-interactional, and propositional complexity, the latter two of which have received relatively less attention in SLA studies. Linguistic complexity can be approached either globally (system complexity) or at the level of local structures (structure complexity). System complexity refers to the linguistic repertoire that learners have in their L2 system. In other words, it involves the range, diversity or variety of different structures that learners use. Structure complexity refers to the depth or sophistication of individual structures, either in a formal or functional sense. Both system and structure complexity can be evaluated at different domains of language: lexis, morphology, syntax and phonology, and sub-domains of each (see Bulté and Housen, 2012, for more discussion of the model.)

In SLA studies, L2 complexity often refers to linguistic complexity, and lexical and syntactic complexity have been studied as two of its major components. While acknowledging the meaning of complexity in a broad sense and the various components of the construct, the present study focuses on syntactic complexity. Syntactic complexity is also called grammatical complexity in the literature. Grammatical complexity is sometimes interpreted in a broader

sense that involves not only syntactic but also morphological and phonological complexity (Bulté & Housen, 2012), but often the two terms (syntactic and grammatical complexity) are used interchangeably as morphological and phonological complexity have rarely been investigated in SLA research.

The following are some definitions of syntactic or grammatical complexity used in previous L2 literature: "progressively more elaborate language;' 'a greater variety of syntactic patterning" (Foster & Skehan, 1996, p.303); "a wide variety of both basic and sophisticated structures are available and can be accessed quickly" (Wolfe-Quintero, Inagaki & Kim, 1998, p.69); "the range of forms that surface in language production and the degree of sophistication of such forms" (Ortega, 2003, p.492). As evident in these definitions, previous researchers have related syntactic complexity to forms of linguistic structures and have understood the construct in terms of (1) range or variety and (2) the degree of elaborateness of those structures. Referring to Bulté and Housen's taxonomy model, the scope of definitions covers syntactic complexity in a formal sense, encompassing both system and structure complexity. The present study is also concerned with complexity in this sense.

1.2 Measures of syntactic complexity in SLA research

Bulté and Housen (2012) proposed that the construct of linguistic complexity be examined at three levels. First, researchers need to establish what the construct is at the theoretical level. Then, researchers can think about how the construct is observable in language performance at the observational level. Finally, they address quantifiable measures of performance at the lowest, operational level.

As mentioned above, the present study concerns grammatical complexity (focusing on syntactic complexity) in both systemic and structural senses, which are observed through

grammatical diversity and sophistication, respectively. In the rest of this section, I review how the construct has been operationalized through quantitative measures in the SLA and L2 writing literature. I begin by introducing measures reviewed in three research syntheses (Bulté & Housen, 2012; Ortega, 2003; Wolfe-Quintero, Inagaki, & Kim, 1998). Wolfe-Quintero et al. comprehensively reviewed measures of grammatical complexity and explored the relationship between the measures and second language development in writing. They examined 32 studies on L2 writing published between 1974 and 1996 and categorized grammatical complexity measures used in these studies into three types: frequencies, ratios, and indices (see Table 1)[1].

Table 1

Wolfe-Quintero et al.'s (1998) inventory of grammatical complexity measures

Frequencies	
Reduced clauses	Preposed adjectives
Dependent clauses	Pronouns
Passives	Articles
Passive sentences	Connectors
Adverbial clauses	Transitional connectors
Adjective clauses	Subordinating connectors
Nominal clauses	Coordinating connectors
Prepositional phrases	

Note. *originally categorized as fluency measures by the authors; # = number

[1] Although the authors classified length-based measures such as a clause, sentence and T-unit length as fluency measures, I have included them as complexity measures, following a more conventional view that length-based measures address syntactic complexity (Ortega, 2003).

Table 1 (cont'd)

Ratios	
Measure	Formula
Clause length (MLC)*	#of words / #of clauses
Sentence length (MLS)*	#of words/ #of sentences
T-unit length (MLT)*	#of words/ #of T-units
T-unit complexity ratio (C/T)	#of clauses/ #of T-units
Sentence complexity ratio (C/S)	#of clauses/ #of sentences
Clauses per error-free T-unit (C/EFT)	#of clauses/ #of error-free T-units
Dependent clauses ratio (DC/C)	#of dependent clauses/ #of clauses
Dependent clauses per T-unit (DC/T)	#of dependent clauses/ #of T-units
Adverbial clauses per T-unit (AdvC/T)	#of adverbial clauses/ #of T-units
Complex T-unit ratio (CT/T)	#of complex T-units/ #of T-units
Sentence coordination ratio (T/S)	#of T-units/ #of sentences
Coordinate clauses per T-unit (CC/T)	#of coordinate clauses/ #of T-units
Coordinate phrases per T-unit (CP/T)	#of phrases with coordinators/ #of T-units
Dependent infinitives per T-unit (DI/T)	#of dependent infinitives/ #of T-units
Complex nominals per T-unit (CN/T)	#of nominals/ #of T-units
Passives per T-unit (P/T)	#of passives/ #of T-units
Passives per clause (P/C)	#of passives/ #of clauses
Passives per sentence (P/S)	#of passives/ #of sentences
Indices	
Measure	Formula
Coordination index	#of independent clause coordination/ #of combined clauses
Complexity formula	Score of weighted structures/ #of sentences
Complexity index	Sum of T-unit scores/ #of T-units

Note. *originally categorized as fluency measures by the authors; # = number

Measures that take the form of frequency simply count the number of specific structures. In ratio measures, the occurrence, frequency, or length of one type of unit is expressed in relation to another type of unit. For example, some may count the number of occurrences of passive structures in a sample essay, while others may count how many times the structure occurs per sentence. The former exemplifies a frequency measure, and the latter represents a ratio measure. Some of the commonly used base units for ratio measures are *clauses, T-units, and sentences*. These base units are defined as follows: Clauses refer to a "structure with a subject and a finite verb" (Hunt, 1965, p.15), and are of various types such as independent clauses, main clauses, adjective, adverbial, and nominal clauses (Cooper, 1976; Hunt, 1965). The last three types are dependent clauses which are "instances of relativization and subordination" (Homburg, 1984, p.92). T-units consist of a main clause and "any subordinate clause or non-clausal structure that is attached to or embedded" (Hunt, 1970, p.189). Lastly, sentences are defined as "a group of words delimited with a punctuation mark" (Wolfe-Quintero et al., 1998, p.84). In index measures, various structures are weighted based on their syntactic complexity. For example, in the *complexity formula* measure, different scores (0, 1, and 2) are assigned to grammatical structures according to their complexity or difficulty. Synthesizing the results of the studies, Wolfe-Quintero and her colleagues concluded that *T-unit length* (MLT), *clause length* (MLC), *T-unit complexity ratio* (C/T), *dependent clause ratio* (DC/C), and *dependent clauses per T-unit* (DC/T) were the best measures for L2 writing development.

Ortega (2003) and Bulté and Housen (2012) performed research syntheses similar to that of Wolfe-Quintero et al. Ortega reviewed 21 cross-sectional studies and five longitudinal studies on college-level L2 writing. More recently, Bulté and Housen (2012) reviewed 40 task-based L2 learning studies (published between 1995 and 2008). Five of the studies (Ellis & Yuan, 2004,

2005; Ishikawa, 2007; Révész, 2008; Storch & Wigglesworth, 2007) investigated learners' performance in written tasks. They classified syntactic complexity measures into overall measures, measures at sentential/ clausal/ phrasal levels, and frequency measures of specific structures. The inventory of measures identified in the two syntheses is presented in Table 2.

Table 2

Inventory of (possible) grammatical complexity measures (adapted from Bulté & Housen, 2012 and Ortega, 2003)

Overall	Mean length of T-unit (MLT)
	Mean length of C-unit
	Mean length of turn
	Mean length of AS-unit
	Mean length of utterance
	Mean length of sentence (MLS)
	S-nodes/ T-unit
	S-nodes/ AS-unit
Sentential—Coordination	Coordinated clauses/ Clauses
	T-units/ Sentences (T/S)
Sentential—Subordination	Clauses/ AS-unit
	Clauses/ c-unit
	Clauses/ T-unit (C/T)
	Dependent clauses/ Clause (DC/C)
	# of subordinated clauses
	Subordinate clauses/ Clauses (SC/C)
	Subordinate clauses/ Dependent clauses (SC/DC)
	Subordinate clauses/ T-unit (SC/T)
	Relative clauses/ T-unit (RC/T)
	Verb phrases/ T-unit (VP/T)

Note. * indicates possible measures that have not been used in the literature. # = number

Table 2 (cont'd)

Subsentential (Clausal + Phrasal)	Mean length of clause (MLC)
	S-nodes/ Clause
Clausal	Syntactic arguments/ Clause*
Phrasal	Dependents/ (noun, verb) phrase*
Other (± syntactic sophistication)	Frequency of passive forms
	Frequency of infinitival phrases
	Frequency of conjoined forms
	Frequency of *Wh*-clauses
	Frequency of imperatives
	Frequency of auxiliaries
	Frequency of comparatives
	Frequency of conditionals

Note. * indicates possible measures that have not been used in the literature. # = number

According to Ortega (2003), the six most frequently used measures were *sentence length* (MLS), MLT, MLC, *T-units per sentence* (T/S), C/T, and DC/C. Among the five studies on L2 writing performance investigated by Bulté and Housen (2012), C/T was the most popular measures, followed by MLT and MLC. Clauses per T-unit (C/T) was employed in four studies, and MLT and MLC were employed in two studies. Other measures used in these studies were the frequency of passive forms and several subordination measures such as DC/C and *subordinate clauses per clause* (SC/S), *per dependent clause* (SC/DC) and *per T-unit* (SC/T). Overall, the studies reviewed in these two research syntheses used predominantly length-based measures and measures of amount of subordination, while uses of other measures were limited. Bulté and Housen (2012) pointed out potential problems related to this trend. First, length-based measures such as MLT and MLS can be elevated in many different ways, for example, through the addition of another clause via coordination or subordination, or another nominal, adjectival,

or adverbial phrase. Therefore, these measures can only capture overall or generic syntactic complexity. Subordination measures also have limitations, but in a different sense. They only tap into complexity at the sentential level and, thus, may not capture the full trajectory of L2 development. In addition, researchers have not accounted for different types of subordination. For example, a noun complement clause followed by a verb (as in "I think that…") and more difficult structures such as an objective relative clause have not been treated separately in the literature. These problems were also identified by Norris and Ortega (2009), who called for the use of more specific measures of coordination and phrasal complexity as well as global measures. They also argued for the use of multiple measures that tap into multiple dimensions of complexity. However, according to Bulté and Housen (2012), few researchers have employed multiple measures in a single study.

To summarize, many of the measures whose validities were confirmed by Wolfe-Quintero et al. (1998) have been popularly used in SLA studies in recent years. These measures include global length-based measures (e.g., MLT, MLS) and measures of subordination (e.g., C/T, DC/C, and DC/T). Some new measures that attend to specific structures such as relative clauses and infinitival clauses have emerged. However, measures are still lacking for the examination of complexity at the clausal and phrasal level.

1.2.1 Review of syntactic complexity measures in L2 writing studies (2009-2016)

The inventory of measures introduced in the previous analyses covers most of the measures employed in L2 writing studies published until 2008. In order to see a more recent trend in the field, I searched for measures of syntactic complexity used in 27 empirical studies on L2 writing published after 2009. The categorization of measures followed the previous reviews

(Bulté and Housen, 2012; Ortega, 2003; Wolfe-Quintero et al., 1998). The results are summarized in Table 3.

This inventory shows that ratio measures are still used more frequently than frequency measures. About half of the studies employed at least one measure of overall complexity such as MLS or MLT (14 studies) and a subordination measure such as C/T, DC/T or DC/C (16 studies). Such a prevalence of ratio measures is understandable considering that frequency measures are affected by text length, which makes them less valid than objective measures, as Wolfe-Quintero et al. (1998) pointed out. Some researchers overcame this disadvantage of frequency measures by using normed frequencies or relative frequencies (e.g., Spoelman & Verspoor, 2010; Verspoor, Schmid, & Xu, 2012). Index measures were rarely used.

One of the noticeable trends was the increased use of specific measures. Amount of coordination was investigated both at the sentential and clausal levels. In addition, many researchers tried to capture complexity at the phrasal level, especially for nominal phrases. For example, Bulté and Housen (2014) and Spoelman and Verspoor (2010) calculated the mean length of noun phrases. Some researchers looked into the occurrence of *complex nominals per T-unit* (CN/T) or *per clause* (CN/C). Crossley and McNamara (2011, 2014) and Guo, Crossley, and McNamara (2013) indirectly calculated the length of nominals in subject positions by measuring the mean number of words before the main verb.

Table 3

Inventory of grammatical complexity measures in L2 writing studies (2009-2016)

	Measure	Study
Frequencies		
Subordination	# of embedded (dependent, subordinate) clauses	Spoelman & Verspoor (2010) Guo, Crossley & McNamara (2013)
	Normalized subordinating conjunctions per 100 words	Vyatkina (2012)
Coordination	Normalized coordinating conjunctions per 100 words	Vyatkina (2012)
Specific structures	# of verb phrases	Crossley & McNamara (2014)
	Part of speech (POS) tags	Guo, Crossley & McNamara (2013)
	Incidence of negation, prepositional phrases, subject relative clauses, *that* verb complements, S-bars, and infinitives	Crossley & McNamara (2014)
	Normed frequencies of 78 grammatical features (e.g., different types of nouns, adjectives, and verbs, etc.)	Asención-Delaney et al. (2011)
	Syntactic similarity (measured by the uniformity and consistency of syntactic constructions in the text, using phrasal and syntactic categories)	Crossley & McNamara (2011, 2014) Guo, Crossley & McNamara (2013) Mazgutova & Kormos (2015)
	Frequencies of modifiers	Vyatkina et al. (2015)
	Distribution of types of sentences (fragment, simple, compound, complex, compound-complex)	Spoelman & Verspoor (2010) Verspoor, Schmid, & Xu. (2012)

Note. # = number

Table 3 (cont'd)

	Measure	Study
Frequencies		
Specific structures	Distribution of types of DC (finite-adverbial, nominal, relative vs. nonfinite)	Verspoor, Schmid, & Xu. (2012)
	Distribution of types of VP constructions (present, tense, past tense, present perfect, etc.)	Verspoor, Schmid, & Xu. (2012)
	Types of NPs	Crossley & McNamara (2014)
Ratios		
Overall	Mean length of sentence (MLS)	Ai & Lu (2013) Bulté & Housen (2014) Lu (2011) Vyatkina (2012) Yoon & Polio (2016)
	Mean length of T-unit (MLT)	Ai & Lu (2013) Bulté & Housen (2014) Danzak (2011) Gyllastad et al. (2014) Lu (2011) Mazgutova & Kormos (2015) Verspoor, Schmid, & Xu (2012) Yoon & Polio (2016)
	Length of production unit	Ai & Lu (2013)
	Mean # of high-level constituents (sentences and embedded sentence constituents) per words in sentences	Crossley & McNamara (2011) Guo, Crossley & McNamara (2013)

Note. # = number

Table 3 (cont'd)

	Measure	Study
Ratios		
Sentential-Subordination & coordination	Ratio of finite verb units per sentence (VP/S)	Vyatkina (2012)
	Clauses per sentence (C/S)	Lu (2011)
	Simple sentence ratio (SSR)	Bulté & Housen (2014)
	Compound sentence ratio (CdSR)	Bulté & Housen (2014)
	Complex sentence ratio (CxSR)	Bulté & Housen (2014)
	Compound-complex sentence ratio (CdCxSR)	Bulté & Housen (2014)
Sentential-Subordination	Clauses per T-unit (C/T)	Benevento & Storch (2011) Larsen-Freeman (2006) Llanes & Munoz (2013) Serrano, Llanes & Tragant (2011) Serrano, Tragant & Llanes (2012) Storch (2009) Lu (2011) Yoon & Polio (2016)
	# of embedded subordinate clauses per T-unit (SC/T): Dependent clauses per T-unit (DC/ T)	Ai & Lu (2013) Danzak (2011) Frear & Bitchener (2015) Gyllstad et al. (2014) Guo, Crossley & McNamara (2013); Storch (2009) Lu (2011) Mazgutova & Kormos (2015) Yoon & Polio (2016)

Note. # = number

Table 3 (cont'd)

	Measure	Study
Ratios		
Sentential-Subordination	Adjectival DC/T	Frear & Bitchener (2015)
	Nominal DC/T	Frear & Bitchener (2015)
	Adverbial DC/T	Frear & Bitchener (2015)
	Complex T-units per T-unit (CT/T)	Lu (2011)
	Dependent clauses per clause (DC/C)	Ai & Lu (2013)
		Lu (2011)
		Yoon & Polio (2016)
	Sentence structure: proportion of subordinate clauses: Subclause ratio (SCR)	Norrby & Hakansson (2007)
		Bulté & Housen (2014)
	Verb phrases (VPs) per T-unit (VP/T)	Lu (2011)
		Yoon & Polio (2016)
Sentential-Coordination	T-units per sentence (T/S)	Ai & Lu (2013)
		Lu (2011)
	Coordinate clause ratio (CCR)	Bulté & Housen (2014)
Subsentential (Clausal +Phrasal)	Words per finite verb-unit	Vyatkina (2012)
	Mean length of finite clause (MLC_{fin})	Bulté & Housen (2014)
	Mean length of clause (MLC)	Ai & Lu (2013)
		Gyllastad et al. (2014)
		Lu (2011)
		Vyatkina (2013)
		Yoon & Polio (2016)
Clausal-Coordination	Coordinate phrases per clause (CP/C)	Ai & Lu (2013)
		Lu (2011)
		Vyatkina (2013)

Note. # = number

Table 3 (cont'd)

	Measure	Study
Ratios		
Clausal-Coordination	Coordinate phrases per T-unit (CP/T)	Ai & Lu (2013) Lu (2011)
Phrasal	Mean length of noun phrase (MLNP)	Bulté & Housen (2014) Spoelman & Verspoor (2010)
	Complex nominals per clause (CN/C)	Ai & Lu (2013) Lu (2011) Vyatkina (2013) Yoon & Polio (2016)
	Complex nominals per T-unit (CN/T)	Ai & Lu (2013) Lu (2011) Yoon & Polio (2016)
	Nonfinite VP per clause	Vyatkina (2012)
	Mean # of words before the main verb	Crossley & McNamara (2011, 2014) Guo, Crossley & McNamara (2013)
	Mean # of complex nominals in subject position	Mazgutova & Kormos (2015)
	# of modifiers per NP	Crossley & McNamara (2014) Guo, Crossley & McNamara (2013) Mazgutova & Kormos, 2015

Note. # = number

In addition, most researchers tended to employ more than one measure in their studies, though some instructed SLA researchers still employed one representative measure of complexity (e.g., Frear & Bitchener, 2015; Mazgutova & Kormos, 2015). Employing more than one measure is a desirable trend because complexity is a multidimensional concept that can only be captured by multiple measures. This effort seems to have been accelerated by advances in technology**.** Many previous studies focused on a small number of measures or analyzed small

amounts of data due to the labor-intensiveness of manual analyses (Lu, 2011). Researchers can now automatically compute a variety of syntactic complexity measures (partially) by using recently-developed tools such as computerized profiling (Long, Fey & Channell, 2008), Coh-Metrix (e.g., Guo, Crossley & McNamara, 2013), D-Level Analyzer (Lu, 2009) or Syntactic Complexity Analyzer (Lu, 2010). For example, several studies reviewed above (Ai & Lu, 2013; Lu, 2011; Yoon & Polio, 2016) used Syntactic Complexity Analyzer to automatically compute a number of syntactic complexity measures that have been popularly used in L2 development studies. However, Norris and Ortega (2009) also cautioned that care should be taken when employing more than one measure due to a potential problem of redundancy. Some measures tap into almost identical characteristics of texts even though they look different. For example, C/T and DC/T measure the same trait. MLT and MLS are also quite similar. These measures are likely to be highly correlated with each other, which in turn may violate assumptions for multivariate statistical analyses (Norris & Ortega, 2009).

Finally, syntactic complexity is often defined in terms of the range and the degree of elaborateness of syntactic structures. Widely used measures of syntactic complexity mostly capture the degree of elaboration (by using length measures and subordination measures) but give less attention to the degree of variation (in other words, diversity), though this dimension has been widely investigated in terms of lexical complexity. Although Norris and Ortega reported signs of researchers' interest in measuring complexity as structural diversity in their research synthesis published in 2009, diversity of syntactic structures seems to remain a relatively infrequent concern compared to other dimensions of complexity. Among the studies published after 2009, I found only seven studies out of 27 in which researchers attended to the dimension of variety (Asención-Delaney et al., 2011; Crossley & McNamara, 2011, 2014; Guo,

Crossley & McNamara, 2013; Spoelman & Verspoor, 2010; Verspoor et al., 2012; Vyatkina et al., 2015). They did so by calculating either frequencies or distributions of various grammatical structures. However, the selection of grammatical structures varied from study to study, and the researchers did not specify the rationale behind the inventory of grammatical structures they investigated; thus, the validity of these measures is still unexplored.

1.3 Syntactic complexity and L2 writing

Complexity measures have been used in SLA research to describe L2 learners' performance and measure their proficiency or progress in language learning (Housen & Kuiken, 2009). In L2 writing research specifically, syntactic complexity has been employed for the following purposes: "(a) to evaluate the effects of a pedagogical intervention on the development of grammar, writing ability, or both; (b) to investigate task-related variation in L2 writing; and (c) to assess differences in L2 texts written by learners across proficiency levels and over time" (Ortega, 2012, p.128). In the following two sections (1.3.1 and 1.3.2), I review studies on L2 writing that included syntactic complexity as a research variable, in accordance with these purposes. The first section reviews studies that investigated influences of external factors on writing performance or ability. The second section reviews the literature on syntactic complexity across proficiency levels or its change over time.

1.3.1 Complexity measures as performance descriptors

Numerous researchers have investigated syntactic complexity as a way to assess the influences of learning conditions on L2 writing proficiency. Their work is mostly in the area of instructed SLA research. In these studies, the purpose of measuring complexity was to scrutinize "how and why language competencies develop for specific learners and target languages, in response to particular tasks, teaching, and other stimuli" (Norris & Ortega, 2009, p. 557). In

other words, researchers were interested in how language learners' performance changes under different learning conditions, and complexity measures were employed as dependent variables to describe their performance. In addition to syntactic complexity, most studies also examined constructs such as accuracy, fluency, and global quality. In most cases, researchers employed one or two measures that represent each construct.

1.3.1.1 Effects of a pedagogical intervention

Some researchers have measured the syntactic complexity of L2 learners' written texts to investigate the effects of an intervention or learning context on their writing skill (e.g., Benevento & Storch, 2011; Casanave, 1994; Ishikawa, 1995; Serrano, Llanes, & Tragant, 2011; Serrano, Tragant, & Llanes, 2012; Shang, 2007; Stockwell & Harrington, 2003; Storch, 2009; Storch & Tapper, 2009). Shang (2007), for example, employed a pretest-posttest design to investigate whether EFL students benefit from practice in writing emails. He found that practice in writing and sending emails improved students' overall sentence complexity as well as grammatical accuracy in subsequent email writing. Storch (2009), also employing a pretest-posttest design, investigated the impact of studying in a L2-medium university over a semester on the writing of students. The results showed that students' writing did not improve significantly in terms of syntactic complexity measured by C/T and DC/C, whereas analytic writing scores significantly increased by the end of the semester. Serrano, Llanes, and Tragant (2011) and Serrano, Tragant, and Llanes (2012) were interested in the effects of studying abroad. In the former study, the authors compared the effects of learning in an international and two domestic (intensive and semi-intensive) contexts. Syntactic complexity was measured by C/T and was not found to be different across the three learning contexts. The latter study tracked the English language development of 14 Spanish learners who studied at a UK university for over a

year. Students' written production was examined three times over the year, and the authors found significant increase in C/T measure over time. Llanes and Munoz (2013) examined the effects of the learning context and its interaction with the age of learners. The authors found significant main effects of learning context and age as well as an interaction effect on syntactic complexity. In all these studies, the authors assumed that an increase in syntactic complexity reflected a learning gain, although several of these works did not find a significant increase in this construct.

1.3.1.2 Effects of task- and genre-related variation

Researchers who were interested in task-based language learning (TBLL) investigated how task variations affected language learners' performance. Some were interested in the relationship between task complexity and learners' writing performance (e.g., Frear & Bitchener, 2015; Ishikawa, 2007; Kormos & Trebits, 2012; Kuiken & Vedder, 2007), while others examined effects of manipulating task conditions such as types of planning (e.g., Ellis & Yuan, 2004 ; Storch & Wigglesworth, 2007). These researchers employed syntactic complexity measures together with fluency and accuracy measures to describe learners' language production. Ellis and Yuan (2004) examined how different task planning conditions influence the language that learners use to perform the task. They found that pre-task planning resulted in greater syntactic variety, while online planning contributed to higher accuracy. Frear and Bitchener (2015) replicated Kuiken and Vedder's (2007) study, which investigated the relationship between cognitive task complexity and linguistic complexity, employing more fine-grained measures of syntactic complexity. They found no significant effect of increased task complexity on the ratio of dependent clauses to T-units (DC/T) as a whole as was found in the Kuiken and Vedder study. However, when dependent clauses of different types were examined

separately, they could see varied effects of increased task complexity on complexity measures. For example, the ratio of adverbial clauses to T-units significantly decreased when task complexity increased, while the ratio of adjectival clauses to T-units remained the same.

Some researchers have studied the effects of different writing-task genres or registers and compared writers' performance in terms of syntactic complexity. For example, Asención-Delaney and Collentine (2011) conducted a multidimensional analysis of a written L2 Spanish corpus in order to investigate how learners' language differs in various types of discourse. They factor analyzed various lexical and grammatical features and found different linguistic complexity measures factored together differently depending on the types of stylistic variations: narrative versus expository. Lu (2011) investigated the effect of genre on syntactic complexity measures in his cross-sectional study. Comparing argumentative and narrative essays written by Chinese learners of English, he found that learners produced more complex structures in argumentative essays than in narrative essays. Yoon and Polio (2016) also examined genre differences in their longitudinal study on ESL students' writing development and found similar results to Lu's. One interesting finding was that genre effects were found on the phrase-level measures but not on clause-level measures.

1.3.1.3 Effects of first language (L1)

Owing to their interest in the influence of first language (L1) on L2 writing, Crossley and McNamara (2011) compared the writings of learners with different L1 backgrounds. Looking at various linguistic features, including syntactic complexity, they found that L2 learners were homogenous and that the differences between the L1 and L2 writings were attributed to limited linguistic resources rather than cultural or L1 differences. Lu and Ai (2015) focused on syntactic

complexity and explored this construct in more depth. They found varied patterns in multiple dimensions of syntactic complexity among learners with different L1 backgrounds.

1.3.2 Complexity measures as indices of development and proficiency

Some researchers have placed syntactic complexity as a primary focus of investigation in their studies. They have attempted to confirm whether syntactic complexity measures stand as valid and reliable indices of second language development or global proficiency in the target language (Lu, 2011). Researchers have investigated how complexity measures change across different proficiency levels (e.g., Lu, 2011) or over time (e.g., Hunt, 1970; Stockwell, 2005; Norrby, 2007). The following sections review these studies.

1.3.2.1 Development in writing over time

I have already introduced some studies in the previous section that investigated changes in learner language over time in specific learning contexts or pedagogical interventions (e.g., Casanave, 1994; Benevento & Storch, 2011; Stockwell & Harrington, 2003; Storch, 2009). Here I have included studies in which the construct of complexity was the primary focus of investigation rather than being employed as a way to measure the influence of external factors.

Some researchers have used relatively large corpus data sets to investigate L2 writing development. Bulté and Housen (2014) focused on short-term development in L2 linguistic complexity (both syntactic and lexical). Analyzing essays written by 45 ESL students in the beginning and at the end of the semester in terms of ten syntactic complexity and three lexical diversity measures, they found that not all measures manifested changes over the course of a semester. Significant gains were evident in the length-based measures (MLS and MLT), clause coordination (compound sentence ratio and coordinate clause ratio) and phrasal elaboration (i.e., *mean length of noun phrase* [MLNP]), but not in subordination measures (i.e., *complex sentence*

ratio, compound-complex sentence ratio, and subclause ratio). The result was contrary to Norris and Ortega's (2009) model of syntactic complexity development, which proposed that syntactic sophistication occurs initially through clausal coordination, is then realized through subordination at the intermediate level, and at a more advanced stage, is achieved predominantly by means of clausal and phrasal elaboration rather than subordination at the sentence level. Crossley and McNamara (2014) used the same corpora as the Bulté and Housen study and conducted a similar study employing a different set of syntactic complexity indices. They used 11 Coh-Metrix indices that measure "syntactic variety, syntactic transformations (e.g., negations and questions), syntactic embeddings, incidence of phrase types, and phrase length" (p.5). They found significant changes in learners' texts over the observed period. The texts contained more noun phrases than verb phrases and a greater number of phrasal modifications at the end of the semester. The syntactic similarity score decreased significantly, which indicated that students used a wider variety of syntactic constructions after a semester of study. Yoon and Polio (2016) used self-compiled corpus data collected every two to three weeks throughout a semester to investigate learner language development over time. They found a statistically significant but weak change in the MLS measure over time. Interesting to note were the interaction effects of genre and time on MLT and on one subordination measure (C/T). MLT was longer and C/T was larger in argumentative essays than in narrative essays at the beginning of the semester. However, the differences between the two genres decreased over time: increases in the measures were found only in narrative essays. Overall, they did not find strong indication of development in terms of syntactic complexity over the course of a semester.

Other researchers included a small number of participants in their studies and focused on their individual trajectories in L2 writing development. Vyatkina (2013) observed two novice

learners of German over a more extended period of time: 19 time points over the course of four semesters. She found that the learners' development of syntactic complexity followed a similar pattern initially, but then the learning paths diverged in the last two semesters. While one learner relied on coordination to lengthen sentences, the other used more complex clausal structures. Based on the results, she argued for the importance of employing both global and specific measures of complexity. Some researchers investigated learner development within the Dynamic Systems Theory framework (Larsen-Freeman, 2006; Spoelman & Verspoor, 2010). These researchers were interested in how constructs of language proficiency interact with each other. They emphasized variability between the learners as well as variation within the learner in the development of these constructs. Larsen-Freeman (2006) observed five Chinese learners of English over six months and investigated how fluency, grammatical complexity, accuracy, and vocabulary complexity emerged and developed in their oral and written performance. She found the individual development trajectories to be very different from one another, while at the same time, the whole group seemed to make progress in general. For example, she found one of the participants focused on lexical complexity throughout the observation period, while others focused more on grammatical complexity. Spoelman and Verspoor (2010) conducted a longitudinal study of a beginning Dutch learner of Finnish. They focused on different complexity measures at the word, phrase, and sentence levels and investigated how these measures developed in relation to one another. To capture dynamic developmental processes, they analyzed the interactions among variables. They found that word complexity and sentence complexity grew together, but NP complexity and sentence complexity developed alternately in a competitive manner.

1.3.2.2 Texts written by learners across proficiency levels

Wolf-Quintero and her colleagues (1998) evaluated the results of studies that investigated the relationship between L2 proficiency levels and syntactic complexity measures. Proficiency levels were mostly defined by school level, program level, or a holistic rating of learner writing performance. The researchers reported that two length-based measures, MLC and MLT, and three measures of subordination, C/T, DC/C and DC/T generally showed a positive linear relationship to proficiency levels. Mixed results were reported for coordination measures such as number of T-units per sentence (T/S). Some studies found that the more frequent use of specific structures such as reduced clauses (Homburg, 1984; Monroe, 1975) or passive sentences (Kameen, 1979) were indications of proficiency levels.

Lu (2011) used a corpus of college-level second language writing at various proficiency levels in evaluating the computation tool he created. He calculated 14 measures using the L2 Syntactic Complexity Analyzer and compared the values across three proficiency levels, which were defined by institutional level. He found that six measures linearly increased along the three proficiency levels. These measures were MLC, MLT, CP/C, CP/T, CN/C and CN/T. Gyllstad, Granfeldt, Bernardini and Kallkvist (2014) also found that some measures discriminate certain levels better than others. They reported that MLC was a better measure for advanced-level writing.

Verspoor, Schmid, and Xu (2012) investigated 64 variables related to constructions, chunks, lexicon, and accuracy in the writings of L2 learners at various proficiency levels in order to search for more reliable indices of written language development. They were interested in which measures can discriminate among proficiency levels, which were predefined by holistic

writing scores. They found that MLT was a medium discriminator and that more dependent clauses were used as the proficiency level increased.

One thing to note in these studies is how proficiency level was operationalized. In Lu's study, naturally occurring groups were used to determine proficiency levels, while Verspoor et al. (2012) assessed writing samples holistically and grouped learners based on the scores. This inconsistency in measures of proficiency makes it hard to compare findings across studies. In addition, although Norris and Ortega (2003) observed that operationalizing proficiency levels in terms of holistic ratings provided more homogenous findings than naturally occurring classes or groups (p.502), cautious interpretation is required when proficiency is measured in this way due to the inherent relationship between quantitative complexity measures and holistic scores.

1.4 Grammatical complexity: L2 assessment

In this section, I examine how grammatical competence has been interpreted and operationalized in assessing L2 writing performance in an attempt to compare the ways grammatical complexity has been viewed in the field of language assessment and SLA. The section also contains a review of studies that employed syntactic complexity measures used in SLA research in investigating testing-related issues.

1.4.1 Assessment of grammar performance

Scholars have viewed language proficiency as a many-faceted skill, and many of them have identified grammar as one distinct component of language competence (e.g., Canale & Swain, 1980; Bachman, 1990). However, the assessment of grammatical knowledge has remained relatively neglected in the language-testing field (Purpura, 2004, p.4). Purpura (2004) made one of the first attempts to investigate comprehensively the construct of grammatical knowledge in the testing context (Zandi, 2014). He proposed a general model of grammar in

which he distinguished between grammatical knowledge, ability, and performance. According to Purpura, grammatical knowledge indicates learners' mental representations of informational structures related to grammatical form and meaning, and grammatical ability incorporates both grammatical knowledge and strategic competence for using the knowledge. It is grammatical ability about which assessors attempt to make inferences in testing. These inferences can be made on the basis of grammatical performance, which is "observable manifestation of grammatical ability" (Purpura, 2004, p.87). Rimmer (2006) identified two measurable dimensions of test-takers' grammar performance: accuracy and range. Accuracy is defined as "control of structures and freedom from error". Range refers to "the variety of grammatical structures that test-takers employ" (p.498), and it concerns the number of different structures and their degree of complexity. Rimmer's notion of range thus incorporated both variety and elaboration in grammatical structures and can be understood as an equivalent concept to syntactic complexity in SLA research.

Table 4

References to syntactic complexity in rating scales for writing

Rubric	Level/Score	Descriptors
TOEFL Independent Writing	5	• displays consistent facility in the use of language, demonstrating syntactic variety, appropriate word choice, and idiomaticity, though it may have minor lexical or grammatical errors
	4	• displays facility in the use of language, demonstrating syntactic variety and range of vocabulary, though it will probably have occasional noticeable minor errors in structure, word form, or use of idiomatic language that do not interfere with meaning

Table 4 (cont'd)

Rubric	Level/Score	Descriptors
TOEFL Independent Writing	3	• may display accurate but limited range of syntactic structures and vocabulary
	2	• an accumulation of errors in sentence structure and/or usage
	1	• serious and frequent errors in sentence structure or usage
IELTS Writing band descriptors: Task 1 (Grammatical range and accuracy)	9	• uses a wide range of structures with full flexibility and accuracy; rare minor errors occur only as 'slips'
	8	• uses a wide range of structures • the majority of sentences are error-free • makes only very occasional errors or inappropriacies
	7	• uses a variety of complex structures • produces frequent error-free sentences • has good control of grammar and punctuation but covers the requirements of the task trends, differences or stages
	6	• uses a mix of simple and complex sentence forms • makes some errors in grammar and punctuation but they rarely reduce communication
	5	• uses only a limited range of structures • attempts complex sentences but these tend to be less accurate than simple sentences • may make frequent grammatical errors and punctuation may be faulty; errors can cause some difficulty for the reader
	4	• uses only a very limited range of structures with only rare use of subordinate clauses • some structures are accurate but errors predominate, and punctuation is often faulty
	3	• attempts sentence forms but errors in grammar and punctuation predominate and distort the meaning
	2	• cannot use sentence forms except in memorized phrases
	1	• cannot use sentence forms at all

Table 4 (cont'd)

Rubric	Level/Score	Descriptors
Jacobs et al.'s ESL Composition Profile (Language Use)	25-22	EXCELLENT TO VERY GOOD: effective complex constructions, □ few errors of agreement, tense, number, word order/function, articles, pronouns, prepositions
	21-18	GOOD TO AVERGAGE: excellent but simple constructions □ minor problems in complex constructions □ several errors of agreement, tense, number, word order/function, articles, pronouns, prepositions *but meaning seldom obscured*
	17-11	FAIR TO POOR: major problems in simple/complex constructions □ frequent errors of negation, agreement, tense, number, word order/function, articles, pronouns, prepositions and/or fragments, run-ons, deletions □ *meaning confused or obscured*
	10-5	VERY POOR: virtually no mastery of sentence construction rules □ dominated by errors □ does not communicate □ OR not enough to evaluate

This notion of grammatical performance is manifested in rating scales that are used to evaluate learners' language performance. Table 4 shows how the construct is illustrated in some widely-used rating scales for assessing the writing performance of L2 learners. For example, in the holistic rating scale used for the TOEFL (https://www.ets.org/toefl) independent writing task, test-takers' language use is evaluated in terms of consistency in using a variety of structures accurately. IELTS (https://www.ielts.org/) uses an analytic rating scale that consists of four subscales: task achievement, coherence and cohesion, lexical resources, and grammatical range and accuracy. According to the descriptors in the grammatical range and accuracy section, use of a wide range of structures and use of complex sentences are indications of advanced proficiency. The last example is the language use section of the ESL Composition Profile created by Jacobs, Hartfiel, Hughey, and Wormuth (1981). The descriptor in this rating scale

also refers to the use of complex versus simple constructions in describing writers' performance. Overall, learners' language use is evaluated in terms of the ability to use a variety of structures and complex sentences accurately in a given writing task. Both diversity and degree of sophistication are addressed in assessing L2 writing performance, while syntactic complexity has been mostly captured by measures of depth or sophistication of structures in SLA studies.

1.4.2 Relationship between syntactic complexity measures and human ratings

Recently, there have been some attempts to link human raters' perceptions of writing quality and linguistic features of texts represented by syntactic complexity measures used in the field of SLA. Crossley and McNamara (2014) investigated the relationship between the indices of syntactic complexity that are sensitive to L2 development and human ratings of language use in L2 writing. They computed various indices using Coh-metix, ran correlation analyses to identify measures that are related to human ratings, and then conducted regression analyses in order to examine whether these indices could be predictive of the subjective ratings. They found that, in addition to the production of all clause types (e.g., matrix, coordinating and embedded clauses), the incidences of infinitives and *that* verb complements were strong predictors of higher ratings of writing quality. An interesting finding was that there was a mismatch between syntactic complexity measures that developed in L2 writing over a semester and those that predicted overall writing quality (as measured by the total writing scores and language use scores). Although the development in learner language over the semester was characterized by more reliance on nominal style and phrasal modifications, raters' judgments of writing quality were not strongly predicted by these features.

Similar results were found by Bulté and Housen (2014). They investigated whether human raters' judgments of writing performance based on an analytic rating scale are related to

syntactic complexity measures. They found *Language Use* scores correlated with most of the syntactic complexity measures they examined: MLS, MLT, *the simple sentence ratio* (SSR), *the complex sentence ratio* (CxSR), *the compound-complex sentence ratio* (CdCxSR), *the subclause ratio* (SCR), MLC, and *mean length of noun phrase* (MNLP). Most of these measures were the ones found to be correlated with the overall writing scores as well. However, these measures were not necessarily development-sensitive. For example, a measure of subordination, CxSR, was significantly correlated with writing quality, but it did not increase over time. Conversely, clausal coordination measures significantly increased over time, while they were not significantly correlated with the subjective ratings of writing quality.

Guo, Crossley, and McNamara (2013) were interested in whether the independent and integrated writing tasks of TOEFL elicit similar performances from L2 writers. They investigated which linguistic features, such as syntactic complexity, predict overall writing scores given by human raters and how much such features predict the scores. Their results did not provide evidence that the syntactic complexity indices they investigated (i.e., number of words before the main verb, number of higher-level constituents per word, number of modifiers per noun phrase, syntactic similarity, and number of embedded clauses) can be predictive of writing scores given by human raters. The authors found a potential reason for the results from the test-takers' proficiency level. TOEFL test-takers are generally assumed to be advanced learners of English, and syntactic complexity indices are not strong discriminators of proficiency among learners at this level, as maintained by Norris and Ortega (2009).

Overall, previous studies have reported mixed results regarding whether syntactic complexity measures have a relationship with subjective ratings by human raters. The results are

far from conclusive, as measures examined varied from study to study. In addition, some commonly used measures in SLA such as DC/C and C/T remain to be investigated.

1.5 Summary

Syntactic (i.e., grammatical) complexity refers to the range and the degree of sophistication of the forms that appear in language production (Ortega, 2003). SLA and L2 writing researchers have employed the construct in order to describe learners' performance and to assess changes in learner language over time or across proficiency levels. Grammatical complexity has also been an important factor in the L2 assessment field. The construct is considered crucial in describing grammatical competence; for example, rating rubrics often utilize the complexity of structures as a descriptor of the writing performance of test takers.

However, how the construct is measured in assessing L2 (writing) performance does not coincide with the ways it is conventionally operationalized in SLA and L2 writing research. As Polio (2001) noted, "the various measures of complexity… indicate that variety does not enter in the equations,… yet the terms complex sentences and variety of structures often appear as part of other components on analytic scales" (p.96). Even after a decade, the degrees of sophistication or elaboration of language structures are used to measure the complexity of writing performance in SLA, while the diversity and complexity of structures used by test-takers are also considered in human raters' subjective evaluations of L2 performance.

Addressing the gap, there have been some recent efforts to attend to the structural diversity dimension of syntactic complexity in SLA studies (Asención-Delaney et al., 2011; Crossley & McNamara, 2011, 2014; Guo, Crossley & McNamara, 2013; Spoelman & Verspoor, 2010; Verspoor et al., 2012; Vyatkina et al., 2015). In addition, some researchers have attempted to link the measures used in L2 writing studies and writing quality by investigating the

relationship between the two. Adding to these previous attempts, the present study aims to fill the gap in the literature by proposing a way to tap into the diversity dimension of syntactic complexity. The following section describes the diversity measures that I am proposing.

1.6 Proposed measures of syntactic complexity to capture syntactic diversity/ variety

In the current study, I propose a way to approach the diversity dimension of syntactic complexity from the verb-argument construction perspective. I test whether diverse use of verb-argument constructions (VACs) can be an indicator of L2 writing proficiency and quality. Verb-argument structures and their contribution to sentence form and meaning have been at the center of many sentence processing models from both theoretical linguistics and psycholinguistics through the years (Becini & Goldberg, 2000). In addition, the syntactic configuration of verbs is known to pose challenges to children in their native language acquisition (Alishahi & Stevenson, 2008) as well as to second language learners (Gries & Wulff, 2009). In addition, there is some empirical evidence to show that syntactic constructions can be predicative of writing proficiency. Hinkel (2003) quantitatively analyzed L1 and L2 academic texts and found that the prevalence of simple constructions such as *be*-copula was characteristic of non-native students' writing. She concluded that non-native students' productive range of grammar was relatively small. Jarvis, Grant, Bikowski, and Ferris (2003) compared the linguistic features of higher-rated and lower-rated ESL compositions and found more frequent use of stative *be* verb constructions in lower-rated compositions and of passive constructions in higher-rated compositions. Therefore, I believe that VACs constitute an appropriate domain of grammar for the study of syntactic proficiency.

As a measure of syntactic diversity, I compute the number of VAC types and the corrected type-token ratio of VAC (VAC CTTR) and used them as two diversity measures in the

current study, following conventions in lexical diversity studies. In examining lexical diversity, many researchers have used the corrected type-token ratio instead of the traditional type-token ratio, as a way to lessen the effect of the length variation of the sample essays. Corrected type-token ratio is calculated by dividing the types of words by the square root of twice the tokens (Carroll, 1964). VAC CTTR in the present study is computed in the same way.

In identifying a set of verb-argument structures, I rely on findings in corpus-informed linguistics: 1) construction grammar and 2) corpus-based descriptive grammar. Linguists with constructionist approaches see knowledge of language as consisting of constructions, which are defined as learned parings of form and meaning at different levels of generality. Words and idioms are constructions, and VACs are constructions at a more abstract level (Goldberg & Suttle, 2010). Goldberg (1995) studied constructions that correspond to basic sentence types, which she believed to reflect basic event types that humans experience. The list of VACs studied by Goldberg (1995) and other researchers in the field (e.g., Becini & Goldberg, 2000; Ellis & Ferrerira-Junior, 2009a; 2009b) are provided in Table 5.

Based on corpus findings, Biber, Johansson, Leech, Conrad, Finegan and Quirk (1999) described major clause patterns that are comparable to VAC types identified by constructionists. Some of these major patterns can be further divided in terms of complementation types. This further classification relies on the work of Quirk, Leech, Sartvik, and Greenbaum (1985) (see Table 6).

Table 5

Verb argument constructions

	Descriptive grammar (Biber et al., 1999)	Construction grammar (Goldberg, 1995, and others)	Example
1	Subject—verb phrase		The sun is shining.
2	Subject—verb phrase—obligatory adverbial	SVPP (Intransitive-motion)	My office is in the next building.
3	Subject—verb phrase—subject predicative	$SVC_{(AP)}$ (Intransitive-resultative)	Your dinner seems ready.
4	Subject—verb phrase—direct object	SVO (Transitive)	That lecture bored me.
5	Subject—verb phrase—prepositional object		He is looking after the dog.
6	Subject—verb phrase—indirect object—direct object	SVOO (Ditransitive)	I must send my parents an anniversary card.
7	Subject—verb phrase—direct object—prepositional object	SVOPP (Dative)	I must send an anniversary card to my parents.
8	Subject—verb phrase—direct object—object predicative	$SVOC_{(AP)}$ (Resultative)	You made him angry.
9	Subject—verb phrase—direct object—obligatory adverbial	SVOPP (Caused-motion)	You can put the dish on the table.
10	Passive	Passive construction	My bicycle is broken.
11	Existential *there*	*there* construction	There are books on the table.
12	Extraposition		It was a good idea to leave early.
13	Cleft		It was my mom who called me.

Table 6

Verb complementation types (Quirk et al., 1985)

	Variants	Example
Copular (SVC & SVA)	SV C (Adjective)	The girl seemed restless.
	SV C (Nominal)	William is my friend.
	SV Adverbial	The kitchen is downstairs.
Monotransitive	SV O (NP) with passive	Tom caught the ball.
	SV O (NP) without passive	Paul lacks confidence.
	SV O (that-clause)	I think that we have met.
	SV O (wh-clause)	Can you guess what she said?
	SV O (wh-infinitive)	I learned how to sail a boat.
	SV O (to-infinitive -S)	We've decided to move house.
	SV O (ing -S)	She enjoys playing squash.
	SV O (to-infinitive + S)	They want us to help.
	SV O (ing + S)	I hate the children quarreling.
Complex transitive (SVOC & SVOA)	SVO C (Adjective)	That music drives me mad.
	SVO C (Nominal)	They named the ship 'Zeus.'
	SVO C (Adverbial)	I left the key at home.
	SVO C (to-infinitive)	They knew him to be a spy.
	SVO C (bare infinitive)	I saw her leave the room.
	SVO C (-ing clause)	I heard someone shouting.
	SVO C (-ed clause)	I got the watch repaired.
	SVO O (NP)	They offered her some food.
	SVO AdvP : Dative	Please say something to us.
	SVO O (that-clause)	They told me that I was ill.
	SVO O (wh-clause)	He asked me what time it was.
	SVO O (wh-infinitive)	Mary showed us what to do.
	SVO O (to-infinitive)	I advised Mark to see a doctor.

The coding for verb-argument structures that appeared in participants' essays was conducted against the above lists. The coding procedure is described in more details in Chapter 2.

CHAPTER 2: THE CURRENT STUDY

2.1 Research questions and hypotheses

The current study was guided by three research questions. The first considers the role of syntactic complexity as an index of L2 proficiency.

(1) Does the syntactic complexity of Korean EFL learners' writing production, as measured by various quantitative complexity measures, function as an indicator of proficiency? In addition, does adding diversity measures increase the predictive power of syntactic complexity in discriminating proficiency levels?

Previous studies have reported mixed results on how syntactic complexity measures are related to different proficiency levels or developmental stages. Some researchers indicated that syntactic complexity measures change according to proficiency levels (e.g., Verspoor, Schmid, & Xu, 2012; Lu, 2011) or over time (e.g., Vyatkina, 2013), while others reported that proficiency or time was not a significant predictor of variation for syntactic complexity features (e.g., Biber, Gray, and Staples, 2014; Yoon & Polio, 2016).

Researchers also reported that some measures are better indicators of proficiency levels or development than others. Lu (2011) tested 14 syntactic complexity measures that are also used in the present study. His results indicated that only six of the measures linearly progressed along three proficiency levels and significantly differentiated between them. These measures were mean length of clause (MLC), mean length of T-unit (MLT), coordinate phrases per clause and per T-unit (CP/C and CP/T), and complex nominal per clause and per T-unit (CN/C and CN/T). In addition, post hoc tests revealed that only one of them significantly differentiated among all three levels. The other measures discriminated between levels two and three only. Gyllstad, Granfeldt, Bernardini and Kallkvist (2014) also found that some measures discriminate

certain levels better than others. They reported that MLC was a better measure for advanced level writings. Building on the previous research on the change of individual complexity indices across proficiency levels, I attempt to investigate whether these measures, individually and as a group, can be predictive of different proficiency levels. I also expect to find that adding diversity measures increase the predictive power.

The second research question targets the link between the L2 syntactic complexity and assessment of L2 writing quality.

(2) How do different syntactic complexity measures relate to subjective ratings of writing quality judged by human raters? Which measure(s) best predict writing quality?

Some researchers found that complexity measures are positively correlated with writing quality judged by human raters (Bulté & Housen, 2014; Crossley & MaNamara, 2014; Kuiken & Vedder, 2014). Bulté and Housen (2014) reported that seven out of ten complexity measures they employed (i.e., mean length of sentence (MLS), MLT, simple sentence ratio (SSR), complex sentence ratio (CxSR), subclauses ratio (SCR), mean length of finite clause (MLC_{fin}) and mean length of noun phrase [MLNP]) showed significant positive correlations with overall writing quality. For measuring the writing quality, the authors used the score of rating scale *Language Use* in addition to the mean total score of five rating scales of an analytic rubric. Crossley and McNamara (2014) found that, in addition to the production of all clause types (e.g., matrix, coordinating and embedded clauses), the incidences of infinitives and *that* verb complements were strong predictors of higher ratings of writing quality. In other words, more diverse syntactic structures were related to higher ratings. In line with these results, I predict that *Language Use* scores be highly correlated with many of the syntactic complexity measures tested, and that measures representing the diversity dimension explain these scores better than

elaboration measures. By testing numerous measures that represent various dimensions of syntactic complexity, including the proposed measures of syntactic diversity, the results of the present study are expected to add findings to the literature.

The third research question investigates how the construct of syntactic complexity is interpreted in the L2 assessment field. As discussed in the previous chapter, SLA researchers use rather objective, quantitative indices to measure the level of complexity, while it is assessed by human raters' evaluation of the criterion in L2 assessment field. In addition to investigating the relationship between the two approaches of measurements with the second research question, I looked into how raters reach their judgments of the level of complexity through rater interviews.

(3) How do raters interpret the notion of syntactic complexity that appears on the *Language Use* scale of a given analytic writing rubric?

To my knowledge, no previous studies have directly asked how raters interpret the descriptors in the *Language Use* scale of an analytic rubric in relation to the notion of syntactic complexity used in SLA. Relevant is research on raters' cognitive process in relation to their response to a rating scale for writing assessment (Barkaoui, 2010; Cumming, Kantor, & Powers, 2002; Knoch, 2009; Lumley, 2002; Winke & Lim, 2015). These researchers investigated raters' decision making process, and many of them found variability in rater scoring behaviors. For example, Lumley (2005) reported that his raters often failed to make decisions based on common interpretation of the scale contents and resorted on different strategies, resulting in variability in rating process. Knoch (2009) also reported that her raters used various coping strategies to deal with difficulty in deciding on a score such as assigning a global score in a holistic rather than an analytic way, or disregarding descriptors. Using an eye-tracker, Winke and Lim found results

similar to Knoch's results: Winke and Lim found that raters use the rating rubric in a systematic (from left to right) way and often disregard descriptors, which suggested that raters use a more holistic approach than the rubric designers perhaps would have surmised. Winke and Lim also suggested that a rather low interrater reliability estimate for the Language Use section of the rubric showed that raters had trouble in interpreting or applying the Language Use section. I predict that in my study raters will have difficulties in interpreting Language Use section of the rubric, and their interpretation of the descriptors and resulting rating processes will vary. I expect the interview data will provide some insights on how the notion of syntactic complexity is interpreted in the L2 assessment field.

2.2 Participants

2.2.1 Korean learners of English

The primary data were collected from Korean learners of English at two different institutional levels. I recruited a total of 187 high school students from four high schools and 203 college students from three universities in South Korea. Their proficiency levels varied and were evaluated by an independent English proficiency test, which is described in detail in the instrument section.

The high school participants were enrolled in four high schools in two provinces in Korea. Three schools, namely schools A, B, and C, were located in Seoul, and the other (school D) was located in Gyeongsang province. School A and B were boys' high schools, and School C was a girls' high school. School D was a co-ed school. The number of students recruited from each school was 17, 21, 89, and 60 (27 male and 33 female students), respectively.

Table 7

Korean participants' demographic and learning background

Institutional level	Gender	Grade	Major	Years studying English	Months in English speaking countries
Secondary (N = 187)	65 male 122 female	3 freshman 184 junior	N/A	10.1 mean 2.38 SD 2 low 16 high	164 None 9 0-6 months 4 7-12 months 2 13-18 months 3 19-24 months 2 2 years 1 2.5 years 1 4 years 1 7 years
College (N = 203)	85 male 118 female	14 freshman 27 sophomore 70 junior 92 senior	47 Social science 44 Engineering 35 Humanities and language 34 Education 29 Science and medical 9 Business 5 Arts	14.86 mean 3.23 SD 5 low 21 high	154 None 21 1-6 months 2 13-18 months 4 19-24 months 1 2.5 years 1 4 years 1 5 years 1 10 years

The college students were enrolled in three universities in different provinces. The majority, 146 participants, was from Seoul National University (70 male students and 76 female students). Twenty-eight participants were recruited from Pusan National University (12 male and 16 female students) and 29 were recruited from Gyeogin National University of Education (4 male and 25 female students). They were pursuing a variety of majors such as science and engineering (e.g., electrical engineering, medical science, and animal science), social science

(e.g., economics, politics, and sociology), humanities and language (e.g., linguistics, history, and English literature), and education. About half of the students were in their senior years, and the other half were mostly juniors followed by sophomores and freshmen (see Table 7 for more demographic information).

2.2.2 Raters

A group of raters also participated in the study. I recruited a total of seven raters at Michigan State University. Four of them were faculty members in English Language Center, and three of them were Ph. D students in Studies in Second Language Program at the university. They were all native speakers of English and had a varied amount of teaching and rating experience (see Table 8). The raters were asked to score the writings of Korean learners based on an analytic rubric, and they participated in a 20-minute interview.

Table 8

Raters' teaching and rating background

	Occupation	ESL/EFL teaching experience	ESL/EFL composition rating experience	Abilities to evaluate ESL/EFL compositions (self-evaluation)
Rater 1	ESL instructor	27 years	Yes	Expert
Rater 2	ESL instructor	35 years	Yes	Expert
Rater 3	ESL instructor	5 years	Yes	Expert
Rater 4	ESL instructor	5 years	Yes	Competent
Rater 5	Ph. D. student	7 years	No	Competent
Rater 6	Ph. D. student	6 years	Yes	Competent
Rater 7	Ph. D. student	4 years	Yes	Novice

2.3 Instruments

2.3.1 Writing tasks

As some previous studies have found genre effects on syntactic complexity measures (e.g., Asención-Delaney & Collentine, 2011), I used writing prompts of different genres to elicit writers' use of a variety of linguistic features. Two genres were used: an argumentative and a narrative writing task. Care was taken to select topics that were suitable for both high school and college student groups. First, I searched for writing prompts that were originally designed as timed writing tasks for young adult learners of English. After the initial selection of potential prompts, I asked two high school teachers in Korea to remove any prompts that were socioculturally unknown or irrelevant for either group of learners and to recommend ones that would be interesting to students.

The argumentative essay prompt was adopted from MSU-CELP exam preparation materials. MSU-CELP is an English language examination developed by the English Language Center at Michigan State University. The exam aims to assess English language ability in four areas, namely writing, listening, reading, and speaking at the C2 level of the Common European Framework of Reference (CEFR). As the test was developed for EFL learners and targets not only adult college learners but also high school learners, it seemed reasonable to adopt one of the writing prompts developed for the exam and administer it to the participants of the current study. I chose several writing prompts from the exam preparation materials that are published online and open to public access (http://www.msu-exams.gr/swift.jsp?CMRCode=1807P3P4S) to create an initial list of possible prompts and selected one. The prompt for the narrative writing task was adopted from Yoon and Polio (2016). Both prompts were edited in an attempt to make the two

prompts comparable in length. The final versions of the prompts used in the study are as follows.

- Argumentative writing task (adapted from MSU-CELP Practice Test 2). Teachers sometimes require students to work together on specific projects. Each student then gets a grade based on the group's success. Some students are quite happy to receive a grade based on the work of the group, while others feel that being graded as part of a group is not fair. What is your opinion about being graded as part of a group? Be sure to support your opinion with examples, reasons, and explanations.
- Narrative writing task (adapted from the MSU Corpus). Think about a particularly good or bad teacher or professor that you had. Tell a story about your experience with that teacher. Be sure to fully develop your story by including relevant examples and specific details.

2.3.2 English proficiency test (C-test)

A C-test, which is a type of cloze test, served as an independent measure of English language proficiency of the Korean participants in this study. The strength of cloze tests lies in their practicality. It is a short, paper-based test that is not constrained by time and space limit, and therefore was chosen as a global proficiency measure in the current study. The main purpose of using the C-test in the current study was to group Korean EFL learners into different proficiency levels.

Cloze tests have been used in language research for a long time. In a cloze test, test-takers are given a passage with a number of deleted single words replaced with blanks and are asked to fill in the blanks. Although the issue of which abilities cloze tests actually measure still remains unresolved (Tremblay, 2011), many researchers found evidence supporting the

reliability of the test (Bachman, 1985; Tremblay, 2011). A C-test was developed in an attempt to resolve difficulties with scoring objectively. Since Raatz and Klein-Braley (1981) introduced a new deletion technique to delete the second half of every second word in a text and coined the term *C-test*, researchers have examined various deletion rates and deletion patterns. For example, Sigott and Kobrel tested different deleting patterns such as deleting 2/3 of the words or leaving the first letter only in an attempt to increase the test difficulty (as cited in Babaii & Ansary, 2001, p.212).

For the present project and another project, I and a colleague developed a 45-item C-test with a first-letter deletion pattern and it can be found in Appendix A. The test consisted of three texts taken from online articles, which was designed to present texts at various comprehensibility levels. The texts were of varied length and structural and lexical complexity. The first sentence of each passage was left intact in order to provide an introduction to the passage. Then, we deleted roughly every 8th word except for the first letter and replaced with blanks. We moved blanks either to the preceding or following words when the same words were deleted repeatedly, or the moved blanks we thought to contribute to the overall quality of the test better than the original ones. We piloted the initial version on three native speakers and six advanced learners of English and then we revised the instrument based on the test takers' responses and opinions. We also pilot-tested the revised version with 13 native speakers of English, and they scored between 76% and 96% (Mean = 84.79, *SD* = 6.34). Their responses were used as acceptable answers.

After the administration of the C-test to 390 Korean participants, the reliability and the discriminability of the test were examined. First, the reliability of the test was estimated using Cronbach's alpha. The Cronbach's alpha value of the entire test was .94, and the values for the

three texts were all above .70, which was interpreted as high. The reliability values are presented in Table 9.

Table 9

C-test reliability

	N of items	Cronbach's Alpha	Cronbach's Alpha based on standardized items
ALL	45	.94	.94
Text 1	11	.75	.76
Text 2	14	.83	.82
Text 3	20	.92	.92

In order to examine the level of difficulty of the test items, item facility (IF) and the item discrimination (ID) were checked. IF is an estimate of item difficulty, and ID is a measure of how well a given item discriminates test-takers with high and low ability (Carr, 2011, pp.269-270). The ID values were all positive, and most of them had a value over .20, which was found acceptable by experts (Nelson, 2000). Three items had an ID score below .20 and were therefore removed from the analysis. The IF value for about half of the items (24 items) was between .30 and .70, which is a normally used target range in practice (Carr, 2011, p.270). Eighteen out of 45 had a value lower than .30, which are interpreted as difficult items, and three items were considered easy (i.e., ID above .70). These easy and difficult items were not removed from the analysis as long as their ID values were above .20, as the purpose of the test was to discriminate participants. Item facility and item discrimination values by item are reported in Appendix B.

2.3.3 Language learning background questionnaire

A questionnaire (adopted from Kim, 2014) asked Korean participants to provide information regarding their gender, year in school, age of first exposure to English, experience

living in English-speaking countries, and their perceived proficiency level in speaking, writing, reading, and listening (see Appendix C and D, Korean translation Appendix E and F).

2.3.4 Rater background questionnaire

At the end of the rating session, I asked raters to fill in a questionnaire asking about their teaching and composition-rating experience, their perceived competency in rating student essays, and their familiarity with any language other than English. I adapted a rater background questionnaire used by Winke and Gass (2013) (See Appendix G).

2.3.5 Rating rubric

Human raters scored writings on an analytic rubric scale developed by Polio (2013) and it can be found in Appendix H. Polio revised an analytic scale adapted from Jacobs et al.(1981), which was based on the evaluation of experienced ESL instructors and targets content, organization, vocabulary, language use, and mechanics. The revised scale consists of the same five components, but the descriptors were changed in accordance with raters' comments and perception of the scale (see Connor-Linton & Polio, 2014, p.4, for more information on the revision process of the scale). Polio (2013) reported that the revised scale was more reliable and valid. Four of the five subscales—Content, Organization, Vocabulary and Language Use—are on a scale from zero to 20, and the mechanics scale ranges from zero to 10, for a total score of 90.

2.4 Procedures

In this section, I report the procedures for each participant group.

2.4.1 Korean learners of English

The main data collection was conducted in Korea in the summer of 2015. Korean EFL students were asked to complete one of the two essay writing tasks. High school students

completed the experiment in a regular English classroom, and their English teachers administered the procedures. First, the students completed a language learning background questionnaire. Then, the teacher distributed the writing prompt and the essay sheet. Half of the students in each class were given the argumentative prompt, and the other half received the narrative prompt. The distribution of the two prompts was random. The writing task lasted for about 30 minutes. Then the teacher collected the essays and gave out the proficiency test. The students filled in the blanks in three passages in increasing order of difficulty (passage 1 to 2 to 3) in 15 minutes. I met college students individually in the library or a café in the campus. They performed the same sequence of tasks as the high school students.

2.4.2 Raters

Writings collected from Korean participants were typed and printed, and any personal information was removed before they were given to the raters. The rating procedure was conducted over the Fall semester in 2015. The raters participated in a norming session, scored 74 essays individually over two weeks after the norming session, and participated in an interview after the completion of rating.

The norming session was guided by one of the raters. He had substantial experience in rating and leading workshops for raters, and volunteered for leading the session. The norming session started with an introduction to the project, and then the raters talked about the rating scale first. The guiding rater read through the descriptors and band levels in the scale briefly, and any questions were resolved through discussion.

After reading through the rating scale together, the raters were given a number of sample essays. They rated one sample essay individually and shared what score they gave to the essay and the rationale behind the rating. They continued the discussion until they reached an

agreement on the scores, and then moved to the next sample essay. In total, they rated and discussed four sample essays. The norming session lasted for 90 minutes.

After the norming session, I gave each rater a packet of argumentative essays. In the packet were 10 essays that were common to all raters, and 27 essays that were unique to each individual rater. A rating scale and a scoring sheet were also included. I asked the raters to finish rating within a week, and everyone returned the packet to me in time. Then I distributed a packet containing the same number (37) of narrative essays to the raters and gave them a week to finish scoring.

Within a week after all the raters had returned the scores for narrative essays, I met them individually for a retrospective interview. During the interview, I asked about the raters' overall procedure of rating, their global impression of the rating rubric, and how they interpreted the descriptors in Language Use section of the scale. The interview lasted for 15 to 20 minutes. The interview employed a semi-structured format, which included a set of interview questions to begin with but deviated from them or added more to pursue the topics arising in the course of the interview (Friedman, 2012). The following are the interview questions that were common to all the interviewees.

- Can you walk me through your overall rating process (and specifically rating of language use)?
- What did you think about the language use section of the rating rubric?
- (Showing sample essays that each rater rated) What were you thinking/ what affected you when you were scoring this essay?
- How did you interpret the wording of the rubric? Could you give me some examples?
    - (errors in) complex structures

- (errors in) morphology
- (frequent use of / minimal use of/ no attempt at) complex sentences
- (excellent/ good/ little/ no) sentence variety

- Do you see differences between *complex sentences* and *complex structures*?

2.5 Data analysis

2.5.1 Quantitative analysis

The essays were evaluated by means of human ratings and by a number of quantitative measures gauging L2 syntactic complexity. For computational analyses of syntactic complexity measures and subjective ratings, essays were typed and saved as individual text files on a computer.

2.5.1.1 Proficiency test

According to Brown (1980), there are various methods that have been developed for the scoring of cloze tests, some of which are exact-answer, acceptable-answer, clozentropy, and multiple-choice (Brown, 1980, p.311). Exact-answer scoring counts the exact words used in the original text as correct, while acceptable-answer scoring counts all contextually acceptable answers as correct. Clozentropy is a refined version of acceptable-answer scoring method. It takes frequency of the acceptable answers in a native speaker pretest into account. In multiple-choice scoring method, test-takers are given a set of alternative answers and asked to choose the correct answer. For the present study, multiple-choice scoring method was not an option as I asked participants to write down answers rather than to choose one from the given options. Among the other options, I chose an acceptable-answer scoring method as it was reported to be a more reliable measure in ESL contexts than an exact-word criterion (Oller, 1972). The

participants' response for each blank was counted as correct when it was contextually and grammatically acceptable and started with a given letter.

Each correct answer received one point. I did not employ a partial-credit scoring, meaning grammatically and contextually unacceptable answers were counted as zero. The original total score was 45, but after checking the item difficulty (ID) values, I discarded three items from the analysis. As a result, the maximum possible score became 42.

2.5.1.2 Subjective ratings

Each essay was given five scores based on the analytic rating scale. The sum of these five scores (i.e., Total score) and the score for Language Use category were used as a rater's judgment of the writing quality.

2.5.1.3 Syntactic complexity: Elaboration measures

I computed 14 syntactic complexity measures using an automated tool developed by Lu (2011), namely, Syntactic Complexity Analyzer. Syntactic Complexity Analyzer is a computational system for automatic analysis of syntactic complexity. The system takes written texts as input and computes 14 measures of syntactic complexity that were selected based on the research syntheses by Wolfe-Quintero et al. (1998) and Ortega (2003). (Refer to Lu (2010) for a detailed description of the system.) These measures consist of; 1) three measures of length of production units (mean length of clause [MLC], mean length of sentence [MLS], and mean length of T-unit [MLT], 2) a sentence complexity ratio (number of clauses per sentence [C/S]), 3) four subordination ratios (T-unit complexity ratio [C/T], complex T-unit ratio [CT/T], dependent clause ratio [DC/C], and dependent clauses per T-unit [DC/T]), 4) three coordination measures (coordinate phrases per clause [CP/C], coordinate phrases per T-unit [CP/T], and sentence coordination ratio [T/S]), and 5) three measures that consider the relationship between

particular structures and larger production units (complex nominals per clause [CN/C], complex nominals per T-unit [CN/T], and verb phrases per T-unit [VP/T]).

2.5.1.4 Syntactic complexity: Diversity measures

In addition to measuring how elaborate structures the essays involve, I also investigated how diverse verb-argument constructions (VACs) were used in essays. The coding procedure for the diversity measures was semi-automated. Two corpus tools were used to identify the verb-argument structures in the data set. First, the essays were part-of-speech (POS) tagged by TagAnt (Anthony, 2014). All essays that were saved as text files were entered into the program and the program tagged every word with POS code. Next, a concordance tool called AntConc (Anthony, 2014) was used to identify the instances of verbs and generate the concordance lines of these verbs. Concordance searches for all verbs resulted in 15,298 hits. I manually filtered the retrieved concordances to keep only the lines containing a main verb. A main verb in the present study was operationalized as a tensed verb in a finite clause. When the tense was marked on an auxiliary verb, the following content verb was viewed as a main verb. Through this process, the instances of auxiliary verbs, gerunds, and *to* or bare infinitive verbs were deleted from the database. Consequently, 9135 hits remained for the analysis.

The concordance lines were exported to an Excel sheet for verb-argument structure coding. The coding procedures were as follows. First, I identified the linear structure of each line using the POS tags. Second, phrase structures were identified by grouping constituents together. Then, verb–argument structure codes were assigned. Based on the summaries of English sentence structures from a corpus-based grammar and construction grammar perspective, I began coding with the distribution of 11 verb-argument structures: (1) verb, (2) verb + obligatory adverbial, (3) verb + subjective predicative, (4) verb + direct object, (5) verb +

prepositional object, (6) verb + indirect object + direct object, (7) verb + direct object + prepositional object, (8) verb + direct object + object predicative, (9) verb + direct object + obligatory adverbial, (10) *passive* construction, and (11) *there* construction. The coding process was iterative, as the set of verb–argument structures identified evolved through repeated coding and grouping. Cleft and extraposition constructions and sub-types of several sentence structures emerged during this procedure. The final coding was conducted against the list of verb–argument constructions identified in previous studies. In the end, a total of 39 verb-argument types were identified (Table 10).

Table 10

Verb-argument structures

	Major types	Sub-types	*Example from data*
1	V (*Intranstive*)		*So I cried.*
2	V + Obligatory adverbial (*Copular*)		*…when I was in high school,*
3	V + Subjective predicative (*Copular*)	V + AdjP	*His class is so interesting that..*
		V + NP	*Although I am a student, …*
		V + CP	*My opinion is that group project is unfair.*
		V + PP	*My opinion is that group project is unfair.*
		V + to infinitive	*Second important thing is to evaluate each other.*
		V + past participle	*…he will not get even punished…*
		V + present participle	*The workload will keep rising.*
4	V + Direct object (*Transitive*)	V + NP	*…and then build up my opinion.*
		V + (*that*) clause	*I believe that it is necessary…*

Note. V = verb; AdjP = adjective phrase; NP = noun phrase; CP = complementizer phrase (i.e., clause); PP = prepositional phrase

Table 10 (cont'd)

	Major types	Sub-types	*Example from data*
4	V + Direct object (*Transitive*)	V + *wh* clause	*Most students didn't care what we are doing.*
		V + *wh* to infinitive	N/A
		V + to infinitive	*My class began to laugh at me.*
		V + present participle	*He didn't give up teaching me.*
		V + [NP + to infinitive]	*He wanted me to know what was wrong.*
		V + [NP + V-ingP]	N/A
		V + "CP"	*She said, "You are right."*
		V + so	N/A
5	V + Prepositional object (*Transitive, prepositional verb*)		*Everyone agreed with it.*
6	V + Indirect object + Direct object (*Ditransitive*)	V + NP + NP	*She brought us some snacks.*
		V + NP + *that* clause	*I've never told him I had been there.*
		V + NP + *wh* clause	*A's mother asked me why I hit him.*
		V + NP + *wh* to infinitive	*He showed us how to live our lives.*
		V + NP + to infinitive	*Lots of professors ask students to work together on their projects.*
		V + NP + "CP"	*I asked myself, "Did I have very big fault?"*
7	V + Direct object + Prepositional object (*Dative*)		*The professor gave lots of articles to students.*
8	V + Direct object + Object predictive (*Complex transitive; resultative*)	V + NP + AdjP	*…and it drive one crazy.*

Note. V = verb; AdjP = adjective phrase; NP = noun phrase; CP = complementizer phrase (i.e., clause); PP = prepositional phrase

Table 10 (cont'd)

	Major types	Sub-types	*Example from data*
8	V + Direct object + Object predictive (*Complex transitive; resultative*)	V + NP + NP	*...so I call them professors.*
		V + NP + Adverbial	*...I regard his students as his younger brother.*
		V + NP + to infinitive	*...he encouraged me to study hard.*
		V + NP + bare infinitve	*That simple word made me cry.*
		V + NP + V-edP	*I had made this machine broken.*
		V+ NP + V-ingP	*I saw some groups having problem..*
9	V + Direct object + Obligatory adverbial (*Caused-motion*)		*I still keep that email in my mail box.*
10	Existential *there*		*There are several reasons.*
11	Passive		*Most of the work is only achieved with multiple people.*
12	Extraposition		*But, it was too difficult to have any question about that.*
13	Cleft		*It is not a teacher but a textbook or US drama that make my English mostly grow.*

Note. V = verb; AdjP = adjective phrase; NP = noun phrase; CP = complementizer phrase (i.e., clause); PP = prepositional phrase

After the coding, the number of each VAC type used by each participant in each essay (i.e., subject), and the corrected type-token ratio (CTTR) were calculated.

2.5.2 Qualitative analysis

Recordings of the rater interviews were analyzed in a qualitative manner. The norming session and seven interviews with each rater were audio-recorded and transcribed. Following Chapelle and Duff (2003)'s and Baralt (2012)'s guidelines, data analysis followed an iterative and cyclical process. First, I started the coding process by transcribing the audio-recorded interview broadly. Then I read the transcript several times and took notes. I segmented the data

and coded each segment, assisted by the program NVivo 9. After that, I grouped related codes together. Through the procedure, I identified themes that are particularly relevant to the issue of syntactic complexity and language use.

2.6 Statistical analysis

Quantitative measures were entered into and analyzed by SPSS version 23. In this section, I report what statistical analyses were used to answer the research questions.

First, I investigated the use of syntactic complexity as an index of L2 written language proficiency. The first research question asked whether the syntactic complexity of Korean EFL learners' writing production, as measured by various quantitative complexity measures, function as an indicator of different proficiency levels. In other words, it asked whether syntactic complexity measures can be used to distinguish between proficiency levels. To answer this question, first, the writings of EFL Korean students were divided into three groups according to their English proficiency test scores. Then I conducted a series of one-way analysis of variance (ANOVA) and a discriminant function analysis (DFA). Through ANOVAs, I investigated whether a significant trend across the three proficiency levels existed for a particular complexity index (Homburg, 1984, p.97). Post hoc analyses were conducted to reveal where the difference occurred, if any. DFA investigates "the extent to which a set of measured variables can distinguish—"discriminate"—between members of different groups or distinct levels of another, nominal or possibly ordinal, variable" (Norris, 2015, p.306). It also provides information regarding which measure best discriminates among groups (Homburg, 1984, p.98). In the present study, DFA was conducted to determine whether the selected sets of syntactic complexity measures can predict proficiency group membership and to find the best predictor. In addition, I compared predictive power of different sets of syntactic complexity measures—

elaboration measures, diversity measures, and the combination of all measures by running three separate discriminant analyses. Cross-validation of the study was done by splitting the data into two halves, running one analysis on one half and running the second analysis with the other. The classification accuracy of the two analyses were compared (Norris, 2015).

The second research question asks whether complexity measures are associated with the quality of writing and subjective ratings of language use. I performed correlations in order to investigate the relationship between each complexity measure and two writing quality scores, Total and Language Use scores. Multiple regression analyses were carried out to investigate the relationship between a group of calculated complexity measures and subjective ratings given by human raters.

CHAPTER 3: RESULTS

In this chapter, I first summarize preliminary results obtained from the English proficiency test and the writing task. Descriptive statistics for C-test scores and subjective ratings given on the student essays are presented. I also report the result of proficiency group division based on the C-test scores and the inter-rater reliability for the subjective ratings given by human raters. In the following sections, the results are organized by research question. I first look at the relationship between syntactic complexity measures and English proficiency. Next, I turn to the relationship between the measures and subjective ratings given by human raters. Lastly, I report the results from the rater interviews on their perceptions regarding the syntactic complexity manifested in the rating scale.

3.1 Preliminary results

3.1.1 English proficiency test (C-test) and proficiency-level placement

The Korean students (N = 390) were divided into three proficiency groups based on the result of the C-test. Those who scored over 60% on the test were considered the advanced-level learners (n = 94). Eighty-one students in this group reported that they had taken one of the three proficiency tests: TOEIC, TOEFL, or Test of English Proficiency developed by Seoul National University (TEPS). The average score (converted on a TOEFL iBT scale based on a TEPS versus TOEFL conversion table [http://www.teps.or.kr/Teps/Public/conversion_table.aspx]) was 108.74 (*SD* = 8.57). Students who scored between 30% and 60% (n = 143) were placed into a high-intermediate group. Ninety of them had an average TOEFL iBT score of 98.24 (*SD* = 1.52). Lastly, students who scored below 30% (n = 153) were placed into a low-intermediate group. Only 13 of them reported an independent proficiency test score, and their average score was 89.77 (*SD* = 8.93). The descriptive statistics for the result of the essay ratings are shown in

Table 11. The mean C-test scores increased as the proficiency levels went up, regardless of the prompt. The results of a factorial ANOVA revealed a statistical main effect of proficiency ($F(2, 384) = 1341.84$, $p < .001$, $\eta_p^2 = .88$). A post-hoc Tukey HSD showed statistical differences between all three proficiency groups. However, neither a main effect of genre ($F(1, 384) = .51$, $p = .48$, $\eta_p^2 = .00$) nor the interaction effect between the proficiency level and genre ($F(2,384) = .02$, $p = .98$, $\eta_p^2 = .00$) was identified. The results show that the distribution of participants across proficiency groups was balanced for each genre.

Table 11

Descriptive statistics: C-test and subjective ratings on the writing task

Proficiency	Prompt	N	C-test		Writing Task			
					Language Use		Total	
			Mean	*SD*	Mean	*SD*	Mean	*SD*
Low-intermediate	Argumentative	74	6.07	*3.37*	7.63	*2.86*	34.57	*12.98*
	Narrative	79	5.71	*3.65*	7.49	*3.04*	34.95	*14.29*
	Total	153	5.88	*3.51*	7.56	*2.95*	34.77	*13.63*
High-intermediate	Argumentative	75	18.64	*3.71*	11.79	*2.35*	53.61	*10.60*
	Narrative	68	18.37	*3.51*	11.93	*2.85*	55.32	*10.83*
	Total	143	18.51	*3.61*	11.86	*2.59*	54.42	*10.70*
Advanced	Argumentative	46	29.78	*3.69*	14.67	*2.39*	67.32	*10.63*
	Narrative	48	29.63	*3.46*	13.96	*2.91*	65.64	*11.53*
	Total	94	29.70	*3.56*	14.30	*2.68*	66.46	*11.07*

3.1.2 Subjective ratings on essays

The essays were rated on an analytic rating scale by seven native speakers of English. Twenty of the 390 essays (ten narrative and ten argumentative essays) were commonly evaluated by all seven raters, and the other essays were scored by either one or two raters. The average

scores of the 390 essays on the two subjective rating scales, Language Use and Total (the sum of five rating scales, including Language Use), were 10.76 (*SD* = 3.88) and 49.61 (*SD* = 17.54), respectively. Mean ratings by genre and proficiency group are presented in Table 11.

In order to check for the inter-rater reliability, I calculated Intra-class correlation coefficients (ICCs) on the scores of the common essays rated by all seven raters. I used a two-way mixed effects model with absolute agreement definition, which assumes that each subject (essay) was rated by two or more raters and that these raters were the only raters participating in the study (Landers, 2015). The average rater ICCs for the Total score and each of the sub-sections of the analytic rubric were found to be high. Table 12 presents the results for the Total and for the Language Use section which are of interest in the current study.

Table 12

Inter-rater reliability (ICCs)

	Intra-class Correlation	95% Confidence Interval	
		Lower Bound	Upper Bound
Language Use	.93***	.87	.97
Total	.96***	.91	.98

Note. *** $p < .001$

3.1.3 Relationship between proficiency test scores and subjective ratings

The relationship between proficiency level as measured by the C-test and writing quality as judged by human raters was also investigated. A Pearson's correlation between the C-test scores and Total scores given on the essays found that the effect size of the correlation was large ($r(390) = .78$, $p < .001$, 95% CI [.74, .82], $R^2 = .61$). This result shows a strong positive linear

relationship between the proficiency test score and the subjective ratings of writing quality, which supports the validity of the C-test as a writing proficiency measure.

3.2 ANOVAs and discriminant function analyses (DFAs): Research question 1

The first research question asks whether syntactic complexity measures can be used to distinguish between proficiency levels. To address this question, I conducted analyses of variance (ANOVAs) that examined differences in each complexity measure across proficiency levels. I also conducted discriminant function analyses (DFAs), in which a group of syntactic complexity measures were used to discriminate between proficiency groups.

3.2.1 ANOVAs

Results from a series of one-way ANOVAs showed that all measures linearly increased across the three proficiency levels, and the mean differences were significant as a function of proficiency level (Table 13). The effect of proficiency was significant on all these quantitative complexity measures with a Bonferroni adjusted alpha level of .003 except for CT/T. The effect sizes were generally large for length-based measures (i.e., MLC, MLS, and MLT), measures of the relationship between particular structures and production units (i.e., CN/C, CN/T, and VP/T) and diversity measures (i.e., VAC Types and VAC CTTR). The effect sizes for coordination measures (i.e., CP/C, CP/T, and T/S) and subordination measures (C/T, DC/C, and DC/T) were small to medium.

Table 13

Proficiency-level effect on syntactic complexity measures (One-way ANOVAs)

	N	Low-intermediate		High-intermediate		Advanced		df	F	Sig. (p)	Effect size (η_p^2)
		M	*SD*	M	*SD*	M	SD				
MLC	387	6.92	*1.27*	7.78	*1.25*	8.78	*1.33*	2	62.05	< .001	.24
MLS	385	11.58	*3.43*	14.09	*3.55*	17.27	*3.83*	2	73.31	< .001	.28
MLT	387	10.56	*2.64*	12.52	*2.98*	15.15	*3.49*	2	68.17	< .001	.26
C/S	386	1.69	*0.50*	1.81	*0.40*	1.99	*0.41*	2	12.43	< .001	.06
C/T	385	1.54	*0.38*	1.60	*0.30*	1.72	*0.33*	2	8.43	< .001	.04
CT/T	390	0.44	*0.22*	0.46	*0.17*	0.52	*0.15*	2	6.07	.003	.03
DC/C	389	0.36	*0.13*	0.37	*0.11*	0.42	*0.10*	2	8.24	< .001	.04
DC/T	384	0.56	*0.30*	0.61	*0.26*	0.76	*0.32*	2	13.09	< .001	.06
CP/C	385	0.11	*0.11*	0.14	*0.10*	0.19	*0.10*	2	16.04	< .001	.08
CP/T	386	0.17	*0.17*	0.22	*0.15*	0.32	*0.18*	2	23.13	< .001	.11
T/S	385	1.09	*0.14*	1.13	*0.14*	1.15	*0.11*	2	6.37	.002	.03
CN/C	390	0.71	*0.29*	0.81	*0.22*	0.96	*0.26*	2	28.54	< .001	.13
CN/T	388	1.11	*0.55*	1.32	*0.46*	1.64	*0.54*	2	3.90	< .001	.14
VP/T	385	1.90	*0.46*	2.10	*0.45*	2.43	*0.51*	2	37.40	< .001	.16
Type	390	6.54	*2.70*	9.76	*2.39*	11.64	*2.93*	2	117.92	< .001	.38
CTTR	390	1.21	*0.31*	1.41	*0.26*	1.48	*0.27*	2	3.84	< .001	.14

Table 14 summarizes the between-level differences in post-hoc Tukey HSD tests. The results showed that differences between the nonadjacent levels (i.e., low-intermediate and advanced) were significant. Differences between the adjacent levels (low-intermediate and high-intermediate, and high-intermediate and advanced) were significant in most of the measures, except for the sentence complexity ratio (C/S), four subordination measures, one coordination measure and one diversity measure. Subordination measures (C/T, CT/T, DC/C, and DC/T) and C/S discriminated between the high-intermediate and advanced levels only. The coordination

measure at a sentence level, T/S, and a diversity measure, VAC CTTR, only discriminated the two lower levels.

Table 14

Post-hoc pairwise comparisons (*p* values) between each proficiency level

		Low-intermediate – High-intermediate	Low-intermediate – Advanced	High-intermediate – Advanced
Length of production units	MLC	< .001	< .001	< .001
	MLS	< .001	< .001	< .001
	MLT	< .001	< .001	< .001
Sentence complexity ratio	C/S	.06	< .001	.01
Subordination	C/T	.26	< .001	.02
	CT/T	.65	.00	.03
	DC/C	.66	< .001	.01
	DC/T	.31	< .001	.00
Coordination	CP/C	.03	< .001	.00
	CP/T	.02	< .001	< .001
	T/S	.03	.00	.54
Particular structures	CN/C	.00	< .001	< .001
	CN/T	.00	< .001	< .001
	VP/T	.00	< .001	< .001
Diversity	VAC Types	< .001	< .001	< .001
	VAC CTTR	< .001	< .001	.10

3.2.2 DFAs

The purpose of the DFAs was to examine whether a group of syntactic complexity measures that have been used popularly in the field of SLA and L2 writing are able to predict

proficiency levels. In addition, I wanted to investigate if the proposed diversity measures would add to the predictive power of syntactic complexity measures.

3.2.2.1 Variable selection

As described earlier, 16 measures consisting of 14 syntactic elaboration measures and two diversity measures were originally computed for participants' essays. Before conducting DFAs, three important assumptions for discriminant analyses were checked following Norris' guidelines (2015). First, univariate normality of distribution was checked for each complexity measure at each proficiency level, and outliers were removed. Second, multicollinearity among predictor variables was checked using a correlation analysis. The bivariate correlations between measures are presented in Table 15. When high correlations between measures ($r > .80$) were identified (Field, 2009), a single predictor variable was selected. Finally, sample size was checked. In the current study, the smallest group sample size was 94 (advanced group). In order to reduce the likelihood of model overfitting, I followed a criterion of 15 observations to 1 predictor. Such a ratio allowed me to include six to seven variables in the analyses. Multivariate outliers were also removed before running analyses.

The final set of predictor variables was selected based on the correlation and ANOVA analyses. In selecting one of the strongly correlated measures, variables with higher effect sizes in ANOVA analyses were primarily selected, following previous studies (e.g., Crossley & McNamara, 2009; Crossley, Salsbury & McNamara, 2011; McNamara, Crossley & McCarthy, 2010). In addition, care was taken to select at least one measure for every dimension of syntactic complexity: length of production units, sentence complexity ratio, subordination, coordination, particular structures and diversity. As a result, seven measures remained for the main analyses

(five elaboration measures and two diversity measures): MLC, DC/T, CP/T, T/S, CN/T, VAC Types and VAC CTTR.

Table 15

Bivariate correlations between syntactic complexity measures

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	MLC	–															
2	MLS	.59	–														
3	MLT	.60	.91	–													
4	C/S	-.05	.74	.63	–												
5	C/T	-.09	.60	.68	.86	–											
6	CT/T	-.08	.52	.57	.74	.82	–										
7	DC/C	.02	.56	.62	.71	.75	.79	–									
8	DC/T	.00	.59	.68	.80	.89	.82	.94	–								
9	CP/C	.57	.28	.29	-.10	-.14	-.09	.01	-.02	–							
10	CP/T	.55	.45	.47	.11	.08	.10	.19	.19	.94	–						
11	T/S	.11	.42	.06	.44	.03	.10	.06	.04	.06	.08	–					
12	CN/C	.67	.53	.63	.14	.18	.20	.27	.28	.31	.35	-.01	–				
13	CN/T	.45	.71	.84	.53	.63	.57	.61	.66	.17	.32	-.01	.85	–			
14	VP/T	.30	.77	.86	.73	.84	.70	.73	.81	.10	.28	.05	.40	.73	–		
15	Type	.26	.29	.28	.16	.11	.08	.12	.16	.16	.17	.15	.09	.13	.27	–	
16	CTTR	.20	.22	.25	.14	.13	.10	.11	.14	.04	.07	.05	.13	.17	.26	.77	–

3.2.2.2 Discriminant function analyses: Elaboration and diversity measures

In order to compare the predictive power of syntactic elaboration and diversity measures in classifying learners into proficiency levels, I conducted three discriminant analyses. Researchers can use a discriminant analysis when they have two or more groups (in this study, the groups are the learners at the three proficiency levels) that theoretically differ on several

interval-level independent variables (in this study the independent variables are syntactic elaboration and diversity measures). Researchers use discriminant analysis to analyze the differences among the groups (on the variables under study) and to provide a way to assign (classify) any one learner into the group with which he or she (based on his or her measures of syntactic elaboration and diversity) most closely resembles (see Klecka, 1980, p. 8). In other words, the analysis looks at whether the group assignment can be predicted by the variables under study.

In the first analysis, five elaboration measures were included in the model. The second model included two diversity measures as predictors. In the third model, all seven complexity measures were entered as predictors. All three analyses identified two discriminant functions, the first function accounting respectively for 98.5%, 99.4%, and 98.4% of the discriminant ability of the variables in each model. Therefore, the first function explains most of the discriminant ability of the predictors in all three models. In the first model, the combined functions (1 and 2) showed a significant discriminating ability (Wilks' lambda = .65, $\chi^2(10, N = 363) = 152.24$, $p < .001$), indicating that the combined predictor variables were able to account for around 35% of the actual variance in proficiency between the three groups. In the second model, an overall statistically significant effect was found for the combined functions (Wilks' lambda = .60, χ^2 (4, N = 363) = 184.81, $p < .001$). In the third model, in which all seven measures were entered, the combined functions were found to be significant (Wilks' lambda = .44, χ^2 (14, N = 363) = 294.91, $p < .001$). The combined functions accounted for around 40% and 56% of the actual variance in proficiency between the three groups in Model 2 and Model 3, respectively.

Table 16

Relationship output for individual predictor variables and functions

Model		Variable	Correlation between discriminant function and predictor variables		Standardized canonical discriminant function coefficients	
			Function		Function	
			1	2	1	2
1	Five elaboration measures	MLC	.82*	-.11	.90	-.30
		CN/T	.65*	.15	.01	-.28
		CP/T	.50	.61*	.01	.77
		T/S	.27	-.53*	.29	-.55
		DC/T	.36	.48*	.50	.52
2	Two diversity measures	VAC Types	.94*	.36	1.32	-.67
		VAC CTTR	.45	.89*	-.53	1.39
3	Seven measures combined	VAC Types	.67*	-.57	1.10	-.33
		MLC	.53*	.35	.49	.37
		CN/T	.42*	.42	.30	-.30
		CP/T	.32	.62*	-.02	.41
		VAC CTTR	.32	-.57*	-.53	-.33
		DC/T	.23	.47*	.14	.69
		T/S	.18	-.19*	.13	-.18

Note. * Largest absolute correlation between each variable and any discriminant function. Variables are ordered by absolute size of correlation within function.

In Table 16, the first column, correlation between discriminant function and predictor variables, presents the relationship between individual indices and each function. In the first model including five measures of syntactic elaboration, MLC, CN/T, and CP/T were found to be highly correlated with Function 1 (> .50). MLC was the strongest marker for the function (r = .82) and contributed the most to separating the proficiency groups. In the second analysis with

two diversity measures, VAC Type was a stronger marker than VAC CTTR for the first function (r = .94). In the third analysis in which all seven measures were entered, MLC and VAC Type were the two variables most highly correlated with the first function.

Table 17

Group centroids

		Level	Function	
			1	2
1	Five elaboration measures	Low-intermediate	-0.77	0.06
		High-intermediate	0.06	-0.11
		Advanced	1.09	0.08
2	Two diversity measures	Low-intermediate	-0.95	-0.03
		High-intermediate	0.24	0.08
		Advanced	1.09	-0.07
3	Seven measures combined	Low-intermediate	-1.26	0.09
		High-intermediate	0.23	-0.18
		Advanced	1.58	0.15

Table 17 shows the group centroids for each function. Group centroids are the means of the discriminant function scores by proficiency group (low-intermediate, high-intermediate, and advanced). The discriminant function scores are derived using the discriminant equations, which are similar to equations obtainable through multiple regression. Individual standardized scores are multiplied by the standardized canonical function coefficients (see Table 16) to compute function scores (Ramos & Liow, 2013). Figures 1 through 3 display these proficiency group centroids and individual cases for each analysis in two dimensions. The horizontal dimension displays Function 1, and the vertical dimension displays Function 2. All three figures illustrate that the first function distinguished between the three groups more clearly than the second

function. However, the horizontal distances between the levels were not found to be equal in the first and the second model. In the first model, the distance between the advanced level and the other two levels was larger than the distance between the two lower groups, indicating that the first function distinguished the advanced level from the other two groups rather well (Figure 1). In the second analysis, the first function distinguished much more strongly between the low-intermediate level and the other two levels (Figure 2).

Figure 1

Cases and group centroids for two discriminant functions: 5 elaboration measures

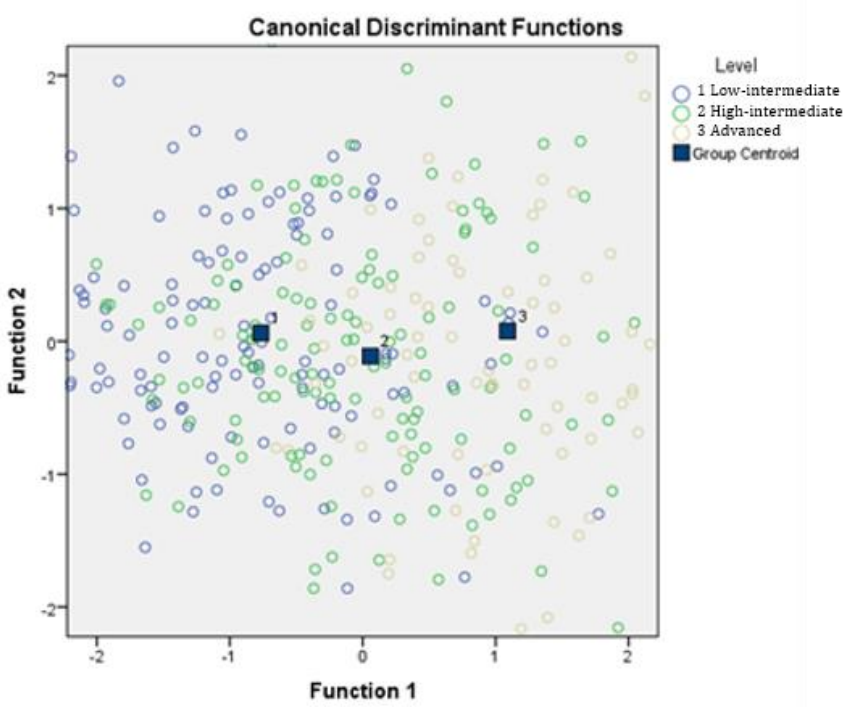
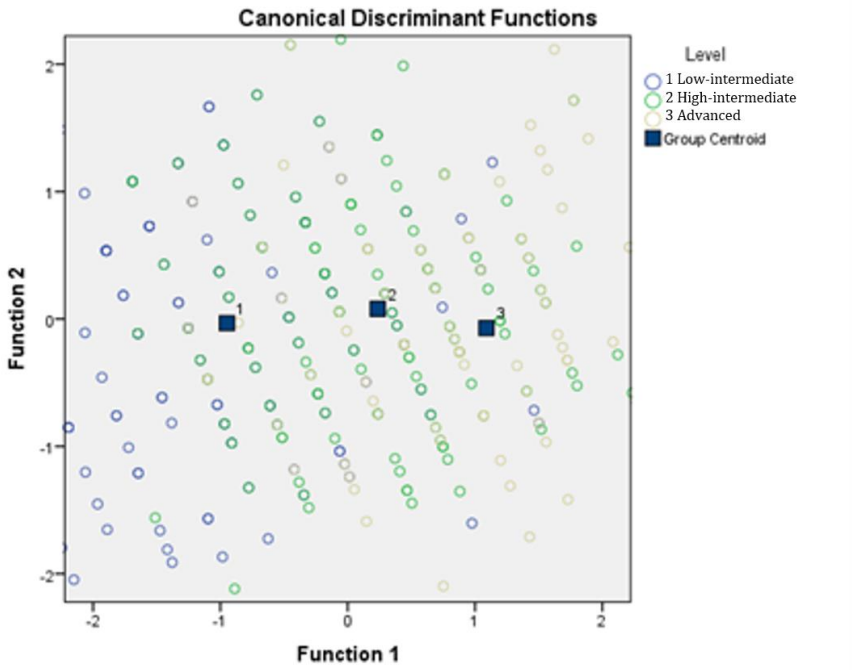


Figure 2

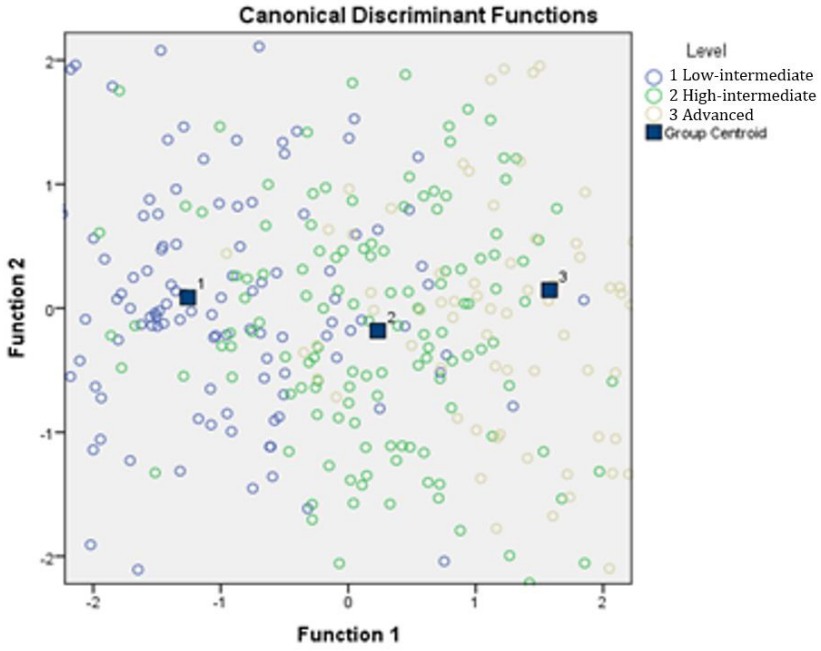
Cases and group centroids for two discriminant functions: 2 diversity measures



Finally, Table 18 shows the classification results for the discriminant analysis. Row sample count presents the predicted frequencies of group (proficiency level) from each analysis, showing the numbers of cases that are correctly and incorrectly classified. For example, 97 cases out of 137 original low-intermediate level cases were correctly predicted, while 34 and 6 cases were predicted to be high-intermediate and advanced level, respectively. Overall, in all three analyses, the combined Functions 1 and 2 were able to classify the cases correctly into the three levels above chance level (i.e., above 33%). The five elaboration measures alone offered a predictive value of 52.9%, and the two diversity measures alone correctly predicted 61.9% of cross-validated grouped cases. The combined use of all seven measures worked best, correctly predicting 68.6% of the cross-validated grouped cases.

Figure 3

Cases and group centroids for two discriminant functions: 7 measures combined



The five elaboration measures were found to be more useful in predicting placement into the low-intermediate level than into the other two groups. Accuracy of predicted placement was much higher for the low-intermediate level, with 67.9% of cross-validated cases predicted correctly. The predictions were substantially less accurate for the high-intermediate and advanced levels, with only 41% and 48% of cases classified accurately. In the second analysis with the two diversity measures, the accuracy of assigning the essays into the low-intermediate level was still superior to the other two levels, with a predictive value of 73%. Compared to the elaboration measures, the diversity measures were extremely useful for the accurate prediction of placement into the high-intermediate level. The prediction was accurate for about 60% of the cases for this level. The prediction accuracy for the advanced group was similar to that of the first model. Lastly, the combined use of all seven measures was found to be the most useful for the accurate prediction of placement. The combined measures exhibited the highest prediction

accuracy in all three levels—the low-intermediate (74%), the high-intermediate (68%), and the advanced level (61%). Although the combined measures still predicted the low-intermediate level essays best, placement into the high-intermediate and advanced levels was more accurate than when either elaboration or diversity measures were used alone. All three levels were classified at a more similar rate compared to the first and the second model.

Table 18

Prediction of group membership according to three discriminant analyses

			Predicted group membership			
			Low-intermediate	High-intermediate	Advanced	Total
Original grouped cases						
Model 1[a]	Raw sample count	Low-intermediate	97	34	6	137
		High-intermediate	51	62	24	137
		Advanced	5	40	44	89
	Percentage	Low-intermediate	**70.80**	24.82	4.38	100.00
		High-intermediate	37.23	**45.26**	17.52	100.00
		Advanced	5.62	44.94	**49.44**	100.00
Model 2[b]	Raw sample count	Low-intermediate	101	33	4	138
		High-intermediate	36	83	20	139
		Advanced	10	37	43	90
	Percentage	Low-intermediate	**73.19**	23.91	2.90	100.00
		High-intermediate	25.90	**59.71**	14.39	100.00
		Advanced	11.11	41.11	**47.78**	100.00
Model 3[c]	Raw sample count	Low-intermediate	103	32	2	137
		High-intermediate	24	94	19	137
		Advanced	1	34	54	89
	Percentage	Low-intermediate	**75.18**	23.36	1.46	100.00
		High-intermediate	17.52	**68.61**	13.87	100.00
		Advanced	1.12	38.20	**60.67**	100.00

Note. [a] 55.9/52.9 of original/cross-validated grouped cases correctly classified; [b] 61.9/61.9% of original/cross-validated grouped cases correctly classified; [c] 69.1/68.6% of original/cross-validated grouped cases correctly classified.

Table 18 (cont'd)

			Predicted group membership			
			Low-intermediate	High-intermediate	Advanced	Total
Cross-validated grouped cases						
Model 1[a]	Raw sample count	Low-intermediate	93	38	6	137
		High-intermediate	53	56	28	137
		Advanced	5	41	43	89
	Percentage	Low-intermediate	**67.88**	27.74	4.38	100.00
		High-intermediate	38.69	**40.88**	20.44	100.00
		Advanced	5.62	46.07	**48.31**	100.00
Model 2[b]	Raw sample count	Low-intermediate	101	33	4	138
		High-intermediate	36	83	20	139
		Advanced	10	37	43	90
	Percentage	Low-intermediate	**73.19**	23.91	2.90	100.00
		High-intermediate	25.90	**59.71**	14.39	100.00
		Advanced	11.11	41.11	**47.78**	100.00
Model 3[c]	Raw sample count	Low-intermediate	102	33	2	137
		High-intermediate	24	93	20	137
		Advanced	1	34	54	89
	Percentage	Low-intermediate	**74.45**	24.09	1.46	100.00
		High-intermediate	17.52	**67.88**	14.60	100.00
		Advanced	1.12	38.20	**60.67**	100.00

Note. [a] 55.9/52.9 of original/cross-validated grouped cases correctly classified; [b] 61.9/61.9% of original/cross-validated grouped cases correctly classified; [c] 69.1/68.6% of original/cross-validated grouped cases correctly classified.

In summary, the results show that complexity measures, either elaboration or diversity measures, or a combination, provided a better overall predictive power for the lowest-level essays than for essays in the two upper proficiency levels. In addition, diversity measures were found to be more useful than elaboration measures in predicting placement into the high-intermediate level. While the five elaboration measures alone predicted the advanced level 7%

better than the upper-intermediate level, the two diversity measures alone exhibited 12% extra predictive value in favor of the high-intermediate level. The use of all seven measures afforded the best predictive placement power into all three levels.

3.3 Correlation and regression analyses: Research question 2

The second research question asks whether syntactic complexity measures are predictive of writing quality as judged by human raters. In order to investigate this question, two correlation analyses were performed between each complexity measure and 1) Language Use scores and 2) Total scores. Two multiple linear regressions were then conducted with each writing quality rating score as an outcome variable and the various syntactic complexity indices as predictor variables.

3.3.1 Correlations

The correlation analyses showed that all syntactic complexity indices were significantly and positively correlated with Total and Language Use scores, indicating that essays with higher scores on these variables tend to be given higher writing scores (Table 19). Differences between the results for Total and Language Use scales were slight. The strongest correlations were found for VAC Types followed by length-based measures—MLT, MLS, and MLC— in both cases. Following Cohen (1992), who defined effect sizes R^2 = .01, .09, and .25 as small, medium, and large effects, respectively, VAC Types was understood to have a strong relationship with writing quality scores. The effect sizes associated with measures for length of production units and particular structures were medium (.09 to .23). The effect sizes for subordination (C/T, CT/T, DC/C) and coordination measures (CP/C, CP/T, T/S) were found to be small (.02 to .09).

Table 19

Correlations between syntactic complexity measures and subjective ratings

		Language Use	Effect size (R^2)	Total	Effect size (R^2)
Length of production units	MLC	.42***	.18	.43***	.18
	MLS	.47***	.22	.47***	.22
	MLT	.48***	.23	.46***	.21
Sentence complexity ratio	C/S	.23***	.05	.22***	.05
Subordination	C/T	.18***	.03	.15**	.02
	CT/T	.17**	.03	.13*	.02
	DC/C	.18***	.03	.15**	.02
	DC/T	.24***	.06	.21***	.04
Coordination	CP/C	.26***	.07	.27***	.07
	CP/T	.29***	.08	.30***	.09
	T/S	.15**	.02	.20***	.04
Particular structures	CN/C	.30***	.09	.29***	.08
	CN/T	.33***	.11	.31***	.10
	VP/T	.37***	.14	.35***	.12
Diversity	VAC Types	.62***	.38	.68***	.46
	VAC CTTR	.37***	.14	.39***	.15

Note. * $p < .05$, ** $p < .01$, *** $p < .001$.

3.3.2 Regression analyses

3.3.2.1 Variable selection

Some assumptions for multivariate analysis were examined before running regression analyses. After controlling for the normality assumption by removing univariate and multivariate outliers and examining the multicollinearity assumption by checking inter-variable correlations, the same set of predictor variables that were used for discriminant function analyses

were selected for the regression analyses: MLC, DC/T, CP/T, T/S, CN/T, VAC Types and VAC CTTR.

The multicollinearity and independent errors assumptions were checked after running the regression analyses. The VIF values were all under 5 (Table 21). The assumption for independent errors was tested with the Durbin-Watson test. The values (= 1.90 for Total score; = 2.00 for Language Use score) were sufficiently close to 2; thus it was assumed that the residuals were uncorrelated.

3.3.2.2 Relationship between syntactic complexity indices and Total score

The standard multiple regression model with seven complexity measures as predictors revealed that there was a statistical relationship between the set of predictor variables and Total score ($F(7, 355) = 73.51$, $p < .001$). The result showed that about 59% of the variance in total scores was accounted for by the set of variables (Table 20).

Table 20

Multiple regression analyses: Model summary

	R	R^2	Adjusted R^2	Std. Error of the Estimate	Durbin-Watson
Language Use	.71	.51	.50	2.63	2.00
Total	.77	.59	.58	10.78	1.90

Table 21

Standard regression coefficients

		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations			Collinearity Statistics	
		B	Std. Error	β			Zero-order	Partial	Part (sr^2)	Tolerance	VIF
Language Use score	(Constant)	-0.32	1.69		-.19	.85					
	MLC	0.65	0.17	.25	3.90	< .001	.47	.20	.15	0.34	2.98
	DC/T	1.19	0.80	.09	1.49	.14	.25	.08	.06	0.36	2.78
	CP/T	0.03	1.04	.00	.02	.98	.32	.00	.00	0.61	1.63
	T/S	1.32	1.15	.04	1.15	.25	.17	.06	.04	0.95	1.05
	CN/T	1.02	0.51	.14	2.00	.05*	.40	.11	.07	0.28	3.60
	VAC Types	0.76	0.07	.66	10.75	< .001	.59	.50	.40	0.37	2.74
	VAC CTTR	-3.11	0.74	-.25	-4.19	< .001	.34	-.22	-.16	0.40	2.51
Total	(Constant)	-1.494	6.91		-0.22	.83					
	MLC	2.37	0.68	.20	3.48	.00**	.47	.18	.12	0.34	2.98
	DC/T	1.60	3.29	.03	0.49	.63	.22	.03	.02	0.36	2.78
	CP/T	2.26	4.28	.02	0.53	.60	.33	.03	.02	0.61	1.63
	T/S	10.58	4.70	.08	2.25	.03*	.22	.12	.08	0.95	1.05
	CN/T	5.58	2.10	.17	2.66	.01***	.38	.14	.09	0.28	3.60
	VAC Types	4.06	0.29	.78	13.96	< .001	.65	.60	.47	0.37	2.74
	VAC CTTR	-17.48	3.04	-.31	-5.75	< .001	.36	-.29	-.19	0.40	2.51

Note. *$p < .05$, ** $p < .01$

The independent relationship between Total score and the predictor variables was examined through regression coefficients and their significance. As shown in Table 21, *t*-test results indicated that MLC, T/S, VAC Types and VAC CTTR contributed uniquely to the outcome variable. DC/T and CP/T did not contribute to the regression model. The relative

importance of each variable was examined by comparing squared semipartial correlations (sr^2) for each term. Among the variables, the strongest predictor of Total score was VAC Types (sr^2 = .47, B = 4.061, β = .783) followed by MLC (sr^2 = .12, B = 2.369, β = .204), CN/T (sr^2 = .09, B = 5.579, β = .171), and T/S (sr^2 = .08, B = 10.583, β = .078).

3.3.2.3 Relationship between syntactic complexity indices and Language Use score

Similar results were found in the regression model that investigated the relationship between syntactic complexity measures and Language Use score. The model was significant ($F(7, 355) = 52.414$, $p < .001$), and the set of variables predicted 51% of the variance in the language use score (Table 20).

Four of the predictor variables independently contributed to Language Use score: MLC, CN/T, VAC Types and VAC CTTR. As was the case for the Total score, VAC Types was the strongest predictor of the language use score (sr^2 = .40, B = 0.762, β = .762). The second and the third strongest predictors were MLC (sr^2 = .15, B = .648, β = .251) and CN/T (sr^2 = .07, B = 1.020, β = .141), respectively.

3.4 Rater interview results: Research question 3

The third research question addresses raters' perceptions of the rating and the rating scale in relation to grammatical complexity. The interviews started with a more general question about their overall rating process and then proceeded with more specific questions regarding rating for the Language Use scale.

3.4.1 Overall rating process

3.4.1.1 Rating sequence

When asked to describe their rating process, most of the raters reported that they started rating by skimming through an essay. Then they rated each subscale of the rubric. Some raters reread (some parts of) the essay to assign the final scores. Which part of the rubric received

attention first varied from rater to rater. Although moving directly from left to right on the rubric—in the order of Content, Organization, Vocabulary, Language Use, and Mechanics—was the most common strategy, not all of the raters suggested that they followed this pattern. One rater reported that he worked in the opposite direction: "For my overall rating, I usually worked backwards on the rubric. I would start with Mechanics, because to me that was sort of the category that was kind of arbitrary in a way. It was a harder one to use" (Rater 5). Rater 1 said that the Language Use section was the area that he addressed first when he was looking at an essay. Another rater reported that the order of scoring differed from essay to essay. He started by rating one or two categories that stuck out after skimming a given essay:

> So for instance, if an essay, regardless of the length, has really good grammatical structure, I might look at the language use category first. And then, other essays that, yeah, the ideas are really strong, if that stands out initially, I'll look at the content band first. So I would say I don't rate in the same order in every essay.
>
> (Rater 4)

3.4.1.2 Lack of information provided by the rating scale

In giving scores for each section of the rubric, a common strategy was to decide in which band to place the essay first and then assign a score within the band:

> I looked at the rubric and kind of decided first within which band I wanted to go. And then usually thinking carefully about the different components of the band depending how high up, you know, between an eleven or a fifteen or whatever, I wanted to give it.
>
> (Rater 3)

> And that was how I would make the decision in terms of what band I was in. And then from there, I would kind of just start considering the strength of each of the qualities. If they're all present, but they're like, "Well, it could be better, I mean, it's there," it'd be a lower score, if it's like, "It's quite good, but it doesn't quite move it to the next band up," then it would still get a score, a higher score, within that band.
>
> (Rater 5)

Some raters noted the difficulty of allocating a specific score within the selected band: "In terms of the Language section, I guess really the most difficult part sometimes was trying to assign scores within a band" (Rater 6). In the next extract, Rater 3 complained:

> In other words it's …with this amount of information I think it is very, very difficult to clearly justify the difference between a seventeen and an eighteen. I think that is very, very arbitrary on the part of the rater. […] it was not difficult to decide most of the time which of the bands it was going to go into, but in going, you know, twelve, thirteen, eleven, it's kind of arbitrary.
>
> (Rater 3)

More experienced raters seemed to have internalized the descriptors in the rubric and had less difficulty in deciding what scores to give: "The rubric, often times it doesn't make any difference as long as I can put them in the proper bands" (Rater 2). Rater 1 is similarly experienced:

> I'm too accustomed to these things. And, really, if you said take out all these descriptors and put these essays in those numbers, I would do that. I don't need that stuff at this point. […] In all language categories, you know, can I sit down and then write a description of why I did that in terms of those kinds of descriptors without looking at that? Yes, I can.
>
> (Rater 1)

Several less experienced raters reported that they referred back to the benchmark essays that all the raters had evaluated during the norming session.

> So, after the norming session, I realized that I tended to rate the higher essays too low and the lower essays too high. So, I was having trouble going to either extreme. So, that was a consideration throughout. So, for the first while, while I still felt like I was getting comfortable, I tended to rate them, and if I thought that it was a poor essay, I would tend to go with my original rating, and then take it down a mark or two, and the same thing for the higher ones, I typically take my marks up if I thought that they were strong essays.
>
> (Rater 5)

3.4.2. Rating process for Language Use

I specifically asked the interviewees to describe their rating process for the language use section of the rubric. I selected three essays that were placed into a high, mid, and low score band by each rater and asked the rater what affected the rating of the essays. The raters read through the essays and reflected on their rating process.

Table 22

Language Use section of the rubric

Score band	Descriptors
20	No major errors in word order or complex structures
	No errors that interfere with comprehension
	Only occasional errors in morphology
	Frequent use of complex sentences
16	Excellent sentence variety
15	Occasional errors in awkward order or complex structures
	Almost no errors that interfere with comprehension
	Attempts, even if not completely successful, at a variety of complex structures
	Some errors in morphology
	Frequent use of complex sentences
11	Good sentence variety
10	Errors in word order or complex structures
	Some errors that interfere with comprehension
	Frequent errors in morphology
	Minimal use of complex sentences
6	Little sentence variety
5	Serious errors in word order or complex structures
	Frequent errors that interfere with comprehension
	Many error in morphology
	Almost no attempt at complex sentences
0	No sentence variety

3.4.2.1 Balancing between accuracy and complexity

In describing the rating process for the language use section, all the raters referred directly to the scale. They seemed to have taken most of the criteria in the rubric into account. One theme that regularly emerged from the interviews was the raters' attempt to balance between two constructs of grammatical ability: accuracy and complexity. I could see that the raters

considered both aspects of grammatical ability while scoring essays for language use, following the descriptors in the rubric. As shown in Table 22, Language Use section of the rubric used in the current study simultaneously addresses both errors and complexity of the language used in writing. For example, an essay is evaluated in terms of the use of complex structures, complex sentences, morphology, and sentence variety and in terms of the errors in them. Rater 4 reflected that he considered both the variety and accuracy of the language:

> So, this essay, to my recollection, has a good variety of grammatical structures. Um, and not only is it **a good variety**, they're **fairly accurate**. […] If I'm looking at the rubric, the top category…so no major errors in word order or complex structures. Looking at this again, I still don't see any global or local errors. No errors that interfere with comprehension. Yeah, I mean, overall, excellent sentence variety. I think there is a pretty good variety of clause structure.
>
> (Rater 4)

Similarly, Rater 6 recalled that, while rating one essay, he was impressed by the use of complex phrases and sentences, and at the same time, he noticed some errors that prevented him from giving a higher score:

> This one … it seems like, I think, I was impressed by the phrasal structures. You know, the essay starts off with, like, "Through the evaluation of the groups' success", kind of a dense, pretty complex phrase to start off with. And I see that throughout. There's also a lot of complex sentences, um, "Although, they put much more efforts to investigate, write a paper, and complete a presentation, it is not fair that some free riders ignoring the parts that should been done can get the same score." There were a few errors here and there, which is probably why I didn't go higher.
>
> (Rater 6)

3.4.2.2 Criteria not specified in the rubric

Several raters mentioned a few factors that affected their rating process although they are not specified in the rubric. One factor that regularly emerged across the interviews was the length or fluency of essays. This issue was often mentioned while the raters were describing their rating process for essays placed into lower bands of the rubric. Raters commented that they gave low scores to short essays because there was not much to evaluate: "I think I want to cross this towards the bottom in part because it was weak. Because there was very little to evaluate" (Rater 1); "You can't say there is? sentence variety, there's really no sentences here" (Rater 5).

Difficulty of grammatical structures was another criterion for rating that was mentioned by several raters. While reflecting on the rating process for a sample high-score essay, Rater 3 referred to good use of a difficult structure in the essay: "Uh, this is very nice, '…which means it could not have been completed…' Nice use of modal, very nice. Modals are hard as I'm sure you know." Rater 4 also thought that his teaching experience and knowledge of structures with which learners' have difficulty affected her rating:

> Use of different adverb phrases. Um…yeah I think a lot of it. I am very heavily influenced by the writing class I teach here that has some pretty explicit grammar objectives. Um, so when I start to see those, a lot of the grammar structures I teach, if I start to see those used accurately in these types of evaluations, that to me is an indicator that these students have flexibility in knowing not only what different subordinators use, for instance. Like, the use of 'so that.' 'So that' uses an adverb of purpose, and 'so' uses a conjunction. That's really tricky for a lot of students. And if I can…if students use those, those are just two of many examples, students are able to use those, pretty accurately, I think that is an indication, and this is obviously a proficiency exam, that to

> me is an indication of proficiency that is higher than um…knowing how to use 'because'. 'Because' is…maybe level two here at the ELC. Yeah, so I think that is mostly what I am looking for: the use, the use of, in the case of adverbs, adverb clauses that are using different subordinating words and phrases with fair amount of accuracy.
>
> (Rater 4)

From his experience, the rater knew that certain subordinators raised more challenge than others, and he gave credit for the use of those structures. He also commented that he did not take into consideration errors in the use of determiners because she knew that to be one of the most challenging aspects in learning English.

3.4.3 Perceptions of the language use section of the rubric

I asked the interviewers how they had perceived the descriptors of the language use section of the rubric. The following themes emerged from their comments.

3.4.3.1 Tension between accuracy and complexity

As described earlier, raters attempted to encompass both accuracy and grammatical complexity in evaluating the language used in essays. However, it was often difficult for them to balance between accuracy and complexity. The following comment from Rater 5 illustrates this challenge:

> The hardest thing for me to balance between was ideas like, so if we look here, attempts, even if not completely successful at a variety of complex structures, right? Uh… but, occasional errors and awkward order complex structures. That one's differentiated here from error in word order or complex structures. So, to me it's kind of difficult in that, ok, they're attempting to use complex structures, which is great,

but let's say that the attempt doesn't work because they don't have the word order correct.

(Rater 5)

Not all the raters valued accuracy and complexity to the same extent. For example, Rater 1 prioritized accuracy over complexity. He recalled that he first considered overall accuracy of the language used in an essay: "…so as soon as I am evaluating an essay, I am looking for accuracy, especially in terms of word order, and phrase construction, and phrase order. […] That's the kind of thing that sticks out more to me" (Rater 1). To Rater 1, accuracy was the construct that gave him the first overall impression of the language used in an essay.

3.4.3.2 Overlap with other categories of the rubric

Another theme that emerged regarding the raters' perceptions of the rubric was an overlap between the language use section and other categories of the rubric. As described earlier, raters often took the length or fluency of the essays into consideration in scoring for the language use category. However, evaluating the "number of words for the amount of time given" is a criterion specified in the content category of the rubric as well. Rater 4 also noted the overlap between the two categories when evaluating the comprehensibility of the language:

> Uh, "Frequent errors that interfere with comprehension." It's completely incomprehensible, right? "(reading a student's essay) Other than he thinks it's fair." What's fair? And that goes to content organization. […] But, I mean, other than that, like, it's just incomprehensible. That starts with the language that you put out there. I mean, you can have perfectly structured English, and not be comprehensible. Right? […] Your language doesn't have to be perfect to be comprehensible, but it at least has to make… combined in some way that one can understand what is

> happening here. Right? Ah, here it just doesn't combine in any way that makes any logical sense. (Rater 5)

In addition, some raters noted that it was often difficult to separate language use from vocabulary use. The distinction between the two categories became problematic especially in relation to morphology issues:

> I think at times it's, it was difficult to like, I'm looking at language usage, and so I'm looking, really to me, language use was basically grammar, in a lot of ways. But, uh, at times, like, morphology comes up with it, so a vocabulary can be tied to morphology to me very easily.
>
> (Rater 5)

> There had been sometimes essays with, either…actually pretty clean in the sense that …very few errors… the syntax, the composition of sentences but some really awkward word choices; um, and then with the, uh, morphology thing: Is that derivational morphology or, you know, or nominalizations… when it should have been an adjective. Things like that, kind of sometimes created a little bit of difficulty to keep those two categories separate.
>
> (Rater 6)

3.4.3.3 Vagueness of descriptors

One of the most frequently mentioned problems that raters faced with the Language Use scale was the vagueness of its descriptors: "I think the, the descriptors… I think that's what's tricky. Language Use is one of the more challenging parts of evaluating essays with this type of rubric" (Rater 4). Difficulty with referring to the rubric due to the vagueness of its descriptors is also reflected in the following comment by Rater 5:

> I think there is always the question, occasional errors in awkward order…well yeah, what is considered a complex structure and what is not considered a complex structure? Um, I, I think my other writing colleagues and I…I think that's…Is there a specific definition of what that means? It almost seems like it's one of those things…we know it when we see it, but we don't know how to define it beforehand.
>
> (Rater 5)

This comment by Rater 5 shows a lack of common and concrete definitions of descriptors that were shared among the raters; thus, the evaluation based on these criteria depended upon the subjective interpretations of raters. Consequently, the raters interpreted the descriptors in the rubric in different ways. In order to detect raters' interpretations of the descriptors, I specifically asked the raters how they understood major terms in the rubric such as morphology, complex structures, complex sentences and sentence variety. Table 23 summarizes their comments.

Morphology was mostly interpreted as word forms or word endings. However, most raters did not comment much on this criterion when describing their rating process. It seemed that they did not pay much attention to the criterion because morphology does not generally raise problems in comprehension, and even the highest score band features essays with (occasional) errors in morphology. The following comment by Rater 1 illustrates this point:

> Morphology. Word endings, which wasn't usually a problem with most of these essays. I would say overall morphology was pretty good. Um, people didn't really have trouble with plurals, which is, you know, one of the easiest parts of, um, English morphology, in my opinion at least. Um, verbs are a little bit harder, but for the most part they weren't that bad.
>
> (Rater 1)

Table 23

Raters' interpretations of the descriptors in the rubric

	Morphology	Complex structures	Complex sentences	Sentence variety
Rater 1	word form	• Multi-clausal sentences • Adverb clauses • Adjective clauses • A properly subordinated clause either placed in front with comma or following a main clause without		• Using simple sentences effectively and then complex sentences effectively • Matrix clause followed by a dependent clause • Dependent clause followed by a matrix clause • Fronting of elements other than a subject • Using simple, compound, and complex sentences, not an overreliance on simple structures or compound structures
Rater 2		• Using clauses (*when*-clauses, adjective clauses, noun clauses, adverb clauses) appropriately • Not just simple sentences, using compound and complex sentences		• Little variety means using the same kinds of sentences all the time.
Rater 3	word endings	• Dependent clauses • Complex noun phrase with pre- or post-modification	• Main clause with dependent clauses	• Clauses led by various subordinators, e.g., *if, even if, wh-words* • Not repeating the same sentence pattern

Table 23 (cont'd)

	Morphology	Complex structures	Complex sentences	Sentence variety
Rater 4		• A variety of adverb clauses • Noun clauses with relative clauses • A fairly well-constructed sentence that has two or three clauses		• Different types of clause structures • Different uses of phrases • Reductions
Rater 5	word form	• Multiple clauses • Relative clauses		
Rater 6	word form	• Passive voice • Perfect aspect • Relative clauses	• Coordinated sentences • Subordinated sentences	
Rater 7		• Transition words • Classic subject + verb + object, choppy and short pattern illustrates a less complex structure		

In most cases, complex structures were interpreted in the same way as they interpreted complex sentences. They were understood to refer to sentences with multiple clauses. Most raters (five out of seven) said they did not distinguish between the two at all. They did not actually notice that the rubric describes complex structures and complex sentences as separate criteria.

> There is some overlap with complex structures and complex sentences and sentence variety. It kind of feels like, if you have one or two of those then you have the rest of them by default.
>
> (Rater 6)

Those who distinguished the two categories did not have the same understanding of the terminology either. One rater mentioned passive voice and perfect aspect as examples of complex structures, while the other referred to complex noun phrases. Lack of reference to other structures does not necessarily mean that they were not considered as complex structures by raters, however. As was evident in an earlier extract from Rater 5, it is possible that raters did not have specific structures in mind until they read and identified particular structure from the essays.

Lastly, to most of the raters, sentence variety meant not using the same sentence pattern repeatedly. What was meant by sentence pattern varied from rater to rater, however. The use of simple, compound, and complex sentences alternately, different clause structures, clauses led by numerous subordinators and varied order of clauses exemplified the notion of sentence variety. The fronting of phrases or clause-to-phrase reduction were also referred to as an indication of diversity in sentence patterns.

CHAPTER 4: DISCUSSION

This chapter summarizes the results and discusses the findings of the current study in relation to previous research in the fields of SLA, L2 writing and L2 assessment. The following sections are organized according to the three research questions.

4.1 Research question 1: Syntactic complexity and proficiency

The first research question for this study asked whether the syntactic complexity of Korean EFL learners' writing production, as measured by various quantitative complexity measures, functions as an indicator of proficiency. The results presented in Chapter 3 show that all syntactic complexity measures examined in the present study linearly increased as the proficiency levels increased. These measures were: 1) three length-based measures (mean length of clause [MLC], mean length of sentence [MLS], and mean length of T-unit [MLT]), 2) a sentence complexity ratio (number of clauses per sentence [C/S]), 3) four subordination ratios (T-unit complexity ratio [C/T], complex T-unit ratio [CT/T], dependent clause ratio [DC/C], and dependent clauses per T-unit [DC/T]), 4) three coordination measures (coordinate phrases per clause [CP/C], coordinate phrases per T-unit [CP/T], and sentence coordination ratio [T/S]), 5) three measures of specific structures (complex nominals per clause [CN/C], complex nominals per T-unit [CN/T], and verb phrases per T-unit [VP/T]), and 6) two diversity measures (verb-argument construction types [VAC Types] and VAC corrected type-token ratio [VAC CTTR]). A series of one-way ANOVAs confirmed that the effect of proficiency level on these complexity measures was significant.

The results for length-based measures were generally consistent with previous studies, in which these measures were confirmed to be an indicator of L2 proficiency. Wolf-Quintero et al. (1998), in their research synthesis, stated that MLC, MLS, and MLT tended to increase at each

proficiency level defined by school or program level, and that these measures discriminated among the levels. Lu (2011) also reported that MLC, MLS, and MLT discriminated the two lowest school levels among four and three nonadjacent levels. Gyllstad et al. (2014) found strong correlations between CEFR scores and MLT and MLC. The results of this study also showed that differences in MLC, MLS, and MLT among all three adjacent proficiency levels were significant. They all increased with a medium effect size ($\geq .05$).

Previous studies have provided mixed opinions regarding subordination measures. Wolfe-Quintero et al. reported that C/T, DC/C, and DC/T increased along with proficiency levels. Gyllstad et al. also found that C/T was a strong discriminator between CEFR levels A (basic) and B (independent) and that the measure increased with respect to the two levels. Lu (2011), however, found that DC/C and DC/T each discriminated nonadjacent levels in a negative rather than progressive direction, and that C/T and CT/T did not discriminate school levels. Biber et al. (2011) also questioned the validity of clausal subordination for assessing grammatical complexity in writing development, arguing that clausal subordination features were characteristics of everyday conversation rather than written language. The results of the current study seem to support Wolfe-Quintero et al.'s findings, in that all subordination measures (i.e., C/T, CT/T, DC/C, and DC/T) linearly increased across the levels, and although the effect sizes were small, the differences due to proficiency were significant for all subordination measures except CT/T. In addition, post-hoc multiple comparison analyses produced results contrary to those of Gyllstad et al. While Gyllstad et al. found C/T to be a discriminator for low proficiency levels, all four subordination measures were found to work as a discriminator for the two higher levels but not for the lower levels in the current study. This result corroborates Norris and Ortega's (2009) claim that subordination measures should be more suitable for measuring

complexity at intermediate and upper-intermediate levels rather than at beginning levels. Norris and Ortega also claimed that these indices are not very useful for measuring advanced levels of L2 development, for which Lu (2011) has provided empirical support. This study has not found evidence for the above claim, however, as the subordination measures continued to increase across the levels. The contradictory results may have been due to differences in the populations investigated in the two studies. Perhaps the advanced level students in this study were not as advanced L2 writers as Lu's highest proficiency group. However, comparison of the average values of each measure in the two studies did not reveal evidence for this difference in levels. The mean values of DC/T for the lowest level was 0.56 in this study, which was comparable to the values for the two lowest levels in Lu's study, 0.52 and 0.54. The measure then decreased to 0.50 and 0.44 in Lu's study, while the mean value continued to increase in the present study (0.61 and 0.76, for the high-intermediate and advanced levels, respectively). The contradictory findings invite more empirical investigations on subordination measures in future research.

The current study reveals a general progression in coordination measures as well. Two phrasal coordination measures, CP/C and CP/T, distinguished among all three levels. A sentential coordination measure, T/S, discriminated the two lower levels, suggesting its greater usefulness as an indicator of lower-level proficiency. The results support Bardovi-Harlig's (1992) proposal that coordination measures can be more sensitive in capturing the proficiency development of beginning learners because coordination can occur at different levels (i.e., phrase, clause, and sentence levels). This study's results suggest that sentential coordination is a more sensitive measure for earlier L2 development.

Three measures of specific structures (CN/C, CN/T, and VP/T) also significantly increased with increases in proficiency level, with a medium effect size, and the measures

distinguished among all three levels. The results for nominal phrases are in line with Lu (2011), who found strong discriminative power in the complex nominals. However, the results for verb phrases (VP/T) were contrary to his findings. This measure also significantly increased along proficiency levels, while Lu did not find significant differences between school levels.

In addition, the results of this study show that indices for syntactic diversity could function as indicators of proficiency. Two diversity measures proposed in the present study were also shown to statistically distinguish all three proficiency levels. VAC Type linearly increased, with a large effect size (≥ 0.25). More advanced L2 learners used more varied types of verb-argument constructions, and the corrected VAC type-token ratios also increased significantly across the proficiency levels. The results are broadly in line with Crossley and McNamara (2014), who found that the syntactic similarity score decreased significantly over time, indicating the association between the production of a wider variety of syntactic constructions and language development.

Researchers have noted the multi-dimensional nature of syntactic complexity and the importance of using multiple measures in assessing the construct, arguing that no single measure is a perfect indicator for proficiency (Norris & Ortega, 2009). In an attempt to add empirical support for this claim, the present study examined whether a number of measures that represent different dimensions of syntactic complexity could, as a group, efficiently function as indicators of writing proficiency. I first investigated the predictive power of the model consisting of five measures of syntactic elaboration in discriminating proficiency levels. The model included one length-of-unit measure (MLC), one subordination measure (DC/T), coordination measures at a phrasal and sentential level (CP/T and T/S), and one measure of nominal phrasal complexity (CN/T). Next, I compared this model with the second model including two measures of

syntactic diversity—VAC Types and VAC CTTR. Finally, I investigated whether the combination of all these measures would increase the predictive power.

Overall, the results of discriminant analyses showed that the sets of syntactic complexity measures could predict proficiency level. In all the models, group memberships were found to be reliably predicted from the set of predictor variables (shown by the significant Wilks's lambda and chi-square results associated with each function). More than half of the cases (essays) were classified correctly by each model (53%, 62%, and 69%, respectively), which was far above chance level (i.e., 33%). These models were most accurate in predicting the low-intermediate level correctly. Accuracy of predicted placement into this level was around 70% in all the models. The diversity measures afforded better predictive power than the elaboration measures overall and were found especially useful in discriminating the low-intermediate and high-intermediate levels. While the elaboration measures incorrectly predicted 39% of the high-intermediate level essays as low-intermediate and predicted 41% correctly, the diversity measures wrongly placed only 26% of the high-intermediate level essays into the low-intermediate level and placed 60% correctly. The third model including all seven measures accounted for the largest variance in proficiency between the three groups and showed the highest predictive accuracy for all three levels, indicating that the proficiency levels can be best predicted by the combination of both elaboration and diversity indices. Among the seven measures, VAC Types was found to contribute the most to the predictive power of the model, followed by MLC and CN/T.

To summarize, the results of the present study suggest that all the investigated syntactic complexity measures can function as indices of proficiency. However, not all the measures functioned in the same manner. Some measures were found to effectively discriminate all levels,

while others discriminate only one adjacent level pair. Overall, ANOVA analyses showed that the diversity measures, length-based measures and measures of particular structures were good indicators of proficiency. They linearly increased along the three proficiency levels, with a medium to large effect size. The subordination and coordination measures functioned better for certain levels than for others. The subordination measures were useful in discriminating the two higher proficiency levels, while a sentential coordination measure, T/S, was more useful in discriminating the two lower levels.

The discriminant analyses suggest that adding the diversity measures to the existing elaboration measures increases the power of the set of complexity measures in predicting proficiency levels. These measures were also weakly correlated with the other 14 elaboration measures (between .04 and .29), which confirms that these measures tap into different aspect of complexity compared with the other measures (Lu, 2011).

4.2 Research question 2: Grammatical complexity and writing quality

The second research question investigated the link between the syntactic complexity measures and the writing quality judged by human raters. Some studies have investigated the relationship between holistic ratings on essays and various complexity indices (e.g., Grant & Ginther, 2000; Taguchi, Crawford, & Wetzel, 2013). Many of these studies used subjective ratings as a way to define proficiency levels. In other words, they placed essays into proficiency levels based on holistic ratings and examined whether complexity indices could function as an index of these levels.

More recently, researchers have more directly investigated the relationship between writing quality perceived by human raters and syntactic complexity indices (Crossley & McNamara, 2014; Guo, Crossley & McNamara, 2013; Kuiken & Vedder, 2014). They

investigated linear relationships between complexity indices and human-judged writing quality. In line with these studies, the current study also investigated the relationship between syntactic complexity indices and writing quality using correlation and regression analyses. For the assessment of writing quality, I used two analytic rating scales, following Crossley and McNamara (2014). I used the sum of five subscales of the analytic rubric (i.e., Content, Organization, Vocabulary, Language Use, and Mechanics) as a judgment of overall writing proficiency and Language Use score as a "more fine-grained measure of syntactic proficiency" (Crossley & McNamara, 2014, p.66).

The results of correlation analyses showed that all 16 syntactic complexity measures, including the two diversity measures newly proposed in this study, were significantly and positively correlated with writing scores given by human raters. The differences in the results for Total and for Language Use scores were slight. The strongest correlations between complexity indices and both writing quality scores were found for a diversity measure, VAC Types, indicating that the use of many different verb-argument structures was related to higher writing quality judged by humans. The second strongest correlations were found for length-based measures (MLC, MLS, and MLT). As was also reported by Bulté and Housen (2014), the longer clausal and sentential units seem to be a sign of higher writing quality.

The correlations between phrasal coordination measures (CP/C, CP/T) and writing quality were moderate, and the correlation between the sentential coordination (T/S) measure and writing quality was found to be weak. The result for T/S is consistent with Bulté and Housen (2014), who found weak and non-significant correlations between writing quality ratings and clausal coordination measures.

Relatively low correlations were found for subordination measures (C/T, CT/T, DC/C, and DC/T), although the relationships were found to be statistically significant. The results for subordination measures differ from Bulté and Housen (2014), in which the authors found a strong relationship between writing quality ratings and a clausal subordination measure. They claimed that more subordination was a characteristic of more advanced writing. However, the results of the current study did not provide clear support for their finding. Kuiken and Vedder (2014) reported that C/T and DC/T were not correlated with raters' judgments of linguistic complexity, which is contrary to Bulté and Housen's findings. As previous studies provided mixed results, more investigation on these measures is needed.

Bulté and Housen (2012) categorized the number of verb phrases per T-unit (VP/T) as a subordination measure as well. Crossley and McNamara (2014) also interpreted the production of fewer verb phrases as indicative of fewer embedded clauses. An interesting observation from the current study was that the relationship between VP/T and writing quality was stronger while very weak correlations were found for other subordination measures, as described above. Somewhat different results found for these measures may have been due to the discrepancy in how the seemingly equivalent measures were defined. Verb phrases identified by the computational tool used in this study, the L2 Syntactic Complexity Analyzer, includes non-finite verb phrases as well as finite verb phrases (Lu, 2010). Therefore, the greater number of verb phrases used per T-unit is not necessarily an indication of more subordination. A closer investigation is needed of the aspect of syntactic complexity addressed by the measure.

Nominal phrasal complexity indices were also found to correlate significantly with writing quality. The number of complex nominals per clause (CN/C) and per T-unit (CN/T) were all positively correlated with Total and Language Use scores. Complex nominals includes

a number of structures: 1) nouns modified by other elements such as adjectives, appositives, or relative clauses, 2) nominal clauses, and 3) gerunds and infinitives in subject position (Cooper, 1976). The result is broadly in line with Crossley and McNamara's (2014) and Bulté and Housen's (2014) findings, although the indices examined in this study and their studies were not identical. Instead of counting the number of complex nominals, they examined a few other comparable indices. Crossley and McNamara examined three indices: average number of modifiers per noun phrase, mean number of words before the main verb, and counted incidence of subject relative clauses. They found a significant positive correlation only for the average number of modifiers per noun phrase. The use of more modifiers in noun phrases was an indication of more proficient writing. Bulté and Housen found a significant correlation between the mean length of noun phrase and writing quality scores, indicating that longer noun phrases are related to higher writing quality. The present study found that more use of complex nominals was related to higher writing quality.

Regression analyses investigated whether a number of selected complexity indices (MLC, DC/T, CP/T, T/S, CN/T, VAC Types and VAC CTTR) would be predictive of writing quality as judged by human raters. The results provided evidence that these measures could significantly predict human judgments of writing quality. The measures jointly accounted for a good proportion of the variance in perceived writing quality. Five measures (MLC, CN/T, T/S, VAC Types and VAC CTTR) were found to be significant contributors to Total scores, accounting for 59% of the variance in the scores. The same measures, with the exception of T/S, were also significant predictors of Language Use scores, accounting for 51% of the variance.

The strongest predictor for both Total and Language Use scores was the number of VAC Types. Essays containing more diverse VACs types were rated higher. Another index of

diversity, VAC CTTR, however, was found to be negatively associated with Total scores when the linear effects of the other variables in the model had been removed. This divergence could have resulted from the relationship between the variable (VAC CTTR) and another predictor variable, VAC Types, given that they are highly correlated (r = .77). In other words, the effect of VAC CTTR may have been subsumed by VAC Types. When the number of VAC Types was held constant, perceived writing proficiency decreased as the corrected type-token ratios increased, which means that fewer verb-argument construction tokens was a sign of lower writing quality. Thus, the measure no longer functioned as a measure of diversity. The result suggests that either of the measures should be selected for multivariate analyses. The next strongest predictor was MLC. Similar to the finding in Bulté and Housen's (2014) study, the results demonstrated that MLC had a large effect on essay scores. The longer clauses were viewed by human raters as an indicator of better writing. The third strongest predictor was CN/T, indicating that the more frequent use of complex nominals was a sign of higher writing quality. The sentential coordination measure, T/S, was the lowest significant predictor for Total scores. This index was not a significant predictor for Language Use scores, however. The phrasal coordination measure (CP/T) and the subordination measure (DC/T) were also not significant predictors of writing quality. Overall, the use of coordination and subordination measures contributed little to human raters' perception of writing quality. This result contradicts Bulté and Housen's (2014) finding that the proportion of simple sentences and the subclause ratio are significant predictors of writing quality.

To combine the findings for Research Questions 1 and 2, the results of the present study demonstrate a similar pattern in the analyses of syntactic complexity measures that are indicators of proficiency and measures that are predictive of human-rated writing quality. First, length-of-

unit measures (MLC, MLS, and MLT) were strong predictors in both analyses. These measures discriminated all three proficiency levels, with large effect sizes, and were significantly correlated with both writing scores rated by human raters. MLC was a strong and significant predictor of writing quality. Second, subordination measures (C/T, CT/T, DC/C, and DC/T) were neither strong predictors of L2 proficiency nor of perceived writing quality. These measures discriminated the high-intermediate and advanced levels significantly, but with small effect sizes. The measures showed low correlations with writing quality ratings. The results for coordination measures showed slightly different patterns. The sentential coordination measure (T/S) discriminated the two lower proficiency levels only, and the correlation between this measure and writing quality was weak. The regression analysis also demonstrated that the measure was not a significant predictor of perceived syntactic proficiency (assessed by Language Use scores). However, it was a significant predictor for overall writing quality (Total scores). Phrasal coordination measures (CP/C and CP/T) distinguished all the levels, with small to medium effect sizes, and were moderately correlated with both writing quality ratings. However, CP/T was not found to be a significant predictor of writing quality. Next, both the number of verb phrases (VP/T) and complex nominals per T-unit (CN/T) were strong discriminators among the three proficiency levels and were moderately correlated with writing quality ratings. The regression analysis confirmed that CN/T was a significant predictor of writing quality. Finally, the two measures of syntactic diversity that were proposed in this study, the number of verb-argument construction types (VAC Types) and the corrected type-token ratio of constructions (VAC CTTR), were strong discriminators of proficiency and also strongly correlated with writing quality ratings. The number of VAC Types was the strongest predictor for both proficiency level and perceived writing quality.

To summarize, the analyses demonstrate that syntactic complexity measures that show significant increase across proficiency levels coincide with the measures that are positively correlated with writing quality ratings in general. The results suggest that syntactic diversity measures, length-of-unit measures, and nominal phrasal indices are good indicators of language proficiency and of writing quality as perceived by human raters. However, the results question the use of subordination or coordination measures as predictors in either case.

4.3 Research question 3: Human raters' perceptions of the Language Use section of an analytic rubric

The third research question asked how human raters interpreted the syntactic (grammatical) complexity descriptors that appear on the Language Use scale of an analytic rating rubric. To answer this question, I performed an individual interview with each rater on their rating process and their interpretation of the descriptors in the rubric.

The interviews started with a general question about the raters' overall rating process. In general, raters read essays first, referred to the rubric to assign scores, and revisited the essays to finalize the scores. This process is consistent with the three-stage model in the rating process described by Lumley (2002) which involves first reading (pre-scoring), then rating for each scoring category, and finally contemplating the given scores.

Previous research has found variability in rating sequence and allocation of attentional focus (Barkaoui, 2010; Cumming, Kantor & Powers, 2002; Lumley, 2002). For example, Cumming, Kantor, and Powers (2002) found that ESL/EFL raters attended to language-related features more extensively than to rhetoric, while native English-speaking raters allotted relatively balanced attention to all main features during holistic rating. Lumley (2005) reported that not all raters rated categories in an orderly way, while Winke and Lim (2015) found that all the raters

assessed categories as arranged in the rubric. Winke and Lim (2015) linked the rating order and the amount of attention given to particular categories to measure of inter-rater reliability. They found that raters paid more attention to the categories that were located in the left part of the rubric, assessing these first, and the authors interpreted the trend as a primacy effect. They also showed how the least attended to category also had the lowest inter-rater reliability (Mechanics). But the results of the current study showed that the order of rating subscales varied depending on the rater, corroborating Lumley's (2005) finding. Although about half of the raters reported that they scored through the rubric from left to right, as in the study by Winke and Lim, several raters reported that they scored a particular section first. One reported that he rated backward on the rubric, starting with Mechanics because it was the most problematic category for him. Another rater reported that the Language Use section was the category that he assessed first, and a third rater described that he started scoring a couple of categories that stood out to him while reading a given essay, resulting in a varied order from essay to essay. The order of rating may or may not reflect the raters' imbalanced attention to different subscales of the rubric. However, the interview data at least did not provide clear evidence that raters weighted any particular category more importantly than others. The raters attended to each subscale of the rubric, and no instance of skipping any subscale was reported. Any indication of a holistic type of rating as was discovered by Knoch (2009) was also not found. All the raters appeared to indicate that they scored each category independently. However, these were self-reported data, and eye-tracking methods, such as those used by Winke and Lim (2015), could verify these self-reported data.

In order to understand raters' perceptions of grammatical complexity manifested in the rubric and its application in scoring, more specific questions were asked on the Language Use

category of the rubric. Raters were asked to describe their scoring procedure for the category and to report their interpretations of the descriptors.

While describing the rating process, raters noted a number of problems with scoring based on the criteria in the rubric. First, balancing between accuracy and complexity was viewed as a difficult task. The Language Use scale considers both accuracy and complexity components of language structures used in texts, and how much to credit incorrect uses of complex structures was up to the raters' discretion. Second, difficulties due to overlaps between the Language Use scale and other categories of the rubric were also reported. Fluency of writing, which is commonly rated for the Content category, also affected the rating of Language Use, and separating Language Use from Vocabulary Use was often considered difficult. Finally, the vagueness of descriptors was one of the most commonly mentioned problems, as was also noted by Knoch (2009). Raters felt that clear definitions of terms were lacking and left up to their subjective interpretations. Consequently, operationalization of some major terms such as 'complex structures' and 'sentence variety' varied from rater to rater. In addition, some descriptors were not interpreted as intended. Most of the raters did not note the distinction between 'complex sentences' and 'complex structures'. Complex structures were often understood to be the same as complex sentences, which are defined as sentences with multiple clauses. Some raters interpreted complex structures as difficult structures, although 'difficulty' was not a criterion explicitly manifested in the rubric. In addition, the raters' teaching experience influenced them when determining difficult structures, which also contributes to the variability among raters.

The notion of grammatical ability in the Language Use category of the rubric used in this study was captured using four separate criteria: complex sentences, complex structures,

morphology, and sentence variety. The raters interpreted these descriptors as follows. First, the raters understood complex sentences as multi-clausal coordinated or subordinated sentences. Linguistic features that were mentioned to exemplify complex structures were complex nominals, passive voice, perfect aspect, transition words, and complex clause patterns. Many of these structures coincided with syntactic complexity indices that have been popularly used in SLA research, such as the number of coordinated phrases or dependent clauses per T-unit, the mean length of noun phrases, or the number of complex nominals. Other examples such as incidences of passive voice and various transition words were reported in Wolfe-Quintero et al.'s (1998) research synthesis, and the distribution of various verb tenses and aspects has been used in some recent studies as well (e.g., Verspoor, Schmid, & Xu; 2012). Next, morphology was understood by most of the raters to indicate word endings. The criterion is intended to be evaluated in terms of accuracy rather than complexity according to the rubric. In other words, the descriptors refer to the number of morphological errors rather than the complexity of morphology used. However, morphological ability seemed to be often disregarded by raters. The criterion was not mentioned as often as other criteria such as complex structures or sentence variety while describing the rating process for the Language Use category. From raters' perspectives, morphology had less room to vary compared with other aspects of grammatical complexity, especially when accuracy of use was considered jointly. Morphology-related errors do not often interfere with understanding, thus raters tended to pay less attention to these features. Lastly, sentence variety was interpreted as the use of diverse types of sentence and clause patterns. How the sentence and clause patterns were interpreted was different from rater to rater. The use of simple, compound, and complex sentences, and clause patterns beyond a simple 'subject-verb-object' pattern were mentioned to exemplify sentence variety.

To summarize, overall, human raters' interpretation of grammatical complexity corresponded to the notion of syntactic complexity common in SLA. Most of the grammatical structures to which raters attended in order to evaluate grammatical complexity coincided with quantitative indices of syntactic complexity used in SLA research. Sentences with multiple clauses via coordination or subordination were the linguistic features most commonly mentioned as exemplifying complex structures. Noun phrases modified by relative clauses or other pre-/ post-modifications were also considered to be an indication of grammatically complex writing. However, the raters did not have identical interpretations of the construct of grammatical complexity. Several major terms in descriptors were abstract and simple, and no further explanations were provided. Consequently, many raters felt that they were not given enough information for scoring from the rubric. They said they needed more specific instructions in order to assign varied scores.

To combine the results for Research Questions 2 and 3, the interview data seem to provide an explanation for the significant correlations between many syntactic complexity measures and writing quality scores reported in the previous section. It is interesting to note, however, that discrepancies were also found between the linguistic features that raters recognized as exemplifications of complex structures and syntactic complexity measures that significantly predicted writing quality. For example, none of the raters directly mentioned clause length in illustrating complex structures, while MLC was a strong predictor of writing quality scores. There are various elements that can lengthen a clause, such as noun phrase modifiers (adjective or prepositional phrases), nonfinite clauses, and adverbial phrases (Norris & Ortega, 2009). Although some raters mentioned noun phrase modifiers, other phrasal modifications were not reported in this study. In addition, although subordinated sentences were the structures that

all raters found complex and regarded as features of advanced writing, subordination measures were not found to be strong predictors of writing quality. Perhaps features that the raters think characterize high-rated essays are not necessarily considered during the rating process. It is also possible that raters could not verbalize all the features they attended to during the rating process. As one rater mentioned, it is possible that raters do not clearly picture what complex structures mean until they find some while reading students' essays. Further research is needed for more direct investigation into raters' cognitive rating process for the language used in L2 learners' essays.

CHAPTER 5: CONCLUSION

In this concluding chapter, I first summarize my research findings in light of the purposes and research questions stated in the beginning of the dissertation. Next, research and practical implications are presented. The chapter concludes with a discussion of the limitations of the current study and suggestions for future research.

5.1 Summary of findings

As outlined in the Chapter 1, this study was designed to accomplish three major purposes. First, the study aimed to determine how syntactic or grammatical complexity has been operationalized in the fields of SLA and L2 assessment and to review the indices of complexity. Second language syntactic complexity in SLA research has often been defined as the degree of sophistication and the range of forms in learner language. Similar to the definition of syntactic complexity in SLA research, in the area of L2 assessment, grammatical performance has been assessed in terms of the number of different structures and their degree of complexity. Through the review of the literature, I found that SLA researchers have noted the multidimensional nature of the construct and developed and used measures that tap into various facets of syntactic complexity. I also found, however, that not all dimensions of the construct have been attributed the same level of importance in research. For example, syntactic complexity at the clausal and phrasal level, and the diversity aspect of complexity have been relatively under-researched. In contrast, both the sophistication and the range of linguistic structures used were main criteria in the assessment of L2 performance.

Noting the gap in the research, the second purpose of this study was to propose measures that address the diversity dimension of syntactic complexity. I proposed to focus on the diverse use of verb-argument constructions. The choice was motivated by the fact that previous studies

in linguistics have found verb-argument constructions to be suitable for evaluating the development of language proficiency. I opted for the number of different verb-argument structure types used and their corrected type-token ratio. Low correlations between the diversity measures and the existing measures of syntactic sophistication demonstrated that the proposed measures tap into an independent trait of complexity.

I also investigated whether the proposed measures, together with traditional complexity measures that have been used in SLA, can be indicative of L2 writing proficiency and writing quality as judged by human raters. The results presented in Chapter 3 showed that most complexity measures worked well as an indicator of proficiency, including the two proposed measures. The complexity measures were also found to be highly correlated with human-rated writing quality. In general, the diversity measures, length of production units, and number of complex noun phrases were better predictors than subordination or coordination measures. The results also showed that adding the diversity measures to the existing elaboration measures increased the predictive power for L2 proficiency and that the number of types of VACs was the strongest predictor of human-rated writing quality. The results lend support to the use of the diversity measures in this area of research.

I also found that notions of grammatical complexity as interpreted by raters overlap with the notion of syntactic complexity in SLA. The data obtained from the rater interviews showed that the raters' conceptualizations of complex structures were comparable to some complexity measures used in SLA, which also explains the high correlations between the measures and the human-rated writing scores. However, variability was found in the interpretations between raters. Another interesting finding was that some features commonly perceived by raters as

characteristic of complex language were not actually found to be strong predictors of writing quality, and some significant predictors were not explicitly recognized by raters.

5.2 Implications

5.2.1 Research implications

The present study has several implications that can inform the fields of SLA, L2 writing, and L2 assessment. First and foremost, the study adds to the literature by proposing measures that capture relatively under-researched aspect of syntactic complexity. The proposed measures are theoretically motivated, and the results of the present study confirm their strong predictive power for L2 proficiency and writing quality as evaluated by human raters. Second, the study provides empirical support for the usefulness of syntactic complexity measures that have been traditionally used in SLA in predicting second language proficiency and second language writing quality. Previous literature has reported conflicting findings on the validity of the measures and pointed out the difficulty of comparing their reliability (e.g., Lu, 2011). The current study overcomes these problems by concurrently examining multiple measures using a large data set. Next, the study also fills a research gap with regard to the link between the understanding of syntactic or grammatical complexity in SLA research and L2 writing assessment practices. The investigation into the relationship between complexity indices used in the field of SLA and writing quality as perceived by human raters offers some insights regarding the link. The data obtained from rater interviews also furthers our understanding of the assessment of grammatical complexity in L2 writing. Finally, the present study investigated writing samples collected from English learners in a non-English speaking country, South Korea. Research on the development of L2 writing has been prevalent in second language contexts (Byrnes, Maxim and Norris, 2010),

and the present study contributes to the understanding of L2 writing development in the instructed FL context.

5.2.2 Practical implications

The results of this study offer several implications that are relevant to practices in L2 assessment. First, the results regarding the relationship between syntactic complexity measures and L2 writing proficiency inform the L2 assessment field by providing insights on what features need to be considered in assessing grammatical ability. Second, the present study reveals some gaps between the factors that affect human judgements of writing quality and human raters' perception of them. The discrepancy requires further investigation, and the results should be reflected in rating scale development/revision and rater training. Next, as described above, this study exposed raters' difficulties in interpreting abstract and vague descriptors in the Language Use section of the analytic rating scale and brought attention to the variability that exists among the raters' operationalizations of major criteria. As Lumley (2002) pointed out, different reactions to a rating scale may result in problems with consistent measurement and interpretation of scores. Efforts to improve raters' common understanding of descriptors are invited. Providing more concrete wording or further illustrations of descriptors would help solve the problem. Rater training would also enhance shared understanding of the rating criteria suitable for a given rating context.

The present study also sheds lights on practices in second language writing pedagogy. The findings regarding the factors that predict L2 proficiency and perceived writing quality can be reflected in syllabus and teaching material development and in teacher education.

5.3 Limitations and future research

The present study has a number of limitations that need to be considered in further research. First, the proficiency test developed and used in this study requires further examination. The use of an independent measure of proficiency was an improvement compared to the previous studies that used school or program levels in that it enables replication and comparisons among studies. However, the test used in the study did not directly assess L2 writing proficiency. Although a positive relationship between the writing scores and the proficiency test scores was confirmed from the data of the current study, more investigation into the validity of the test is recommended.

Second, the present study investigated essays collected from learners with the same first language (L1) background. Although the use of a homogeneous group has the advantage of controlling for variation due to the learners' L1, further research is needed to generalize the findings of the present study to other L1 groups.

Third, the way raters' perceptions of the rating rubric were investigated bares some methodological limitations. As many researchers have stated, verbal reports cannot completely reveal a person's cognitive processes. Features that they implicitly and intuitively attended to may not have been successfully retrieved. In addition, the features mentioned by the raters may not have actually been attended to while rating as the interviews were not concurrently conducted. More data needs to be cumulated on this issue via various data collection methods, such as eye-tracking methods (as done by Winke & Lim, 2015), think-aloud protocols, stimulated-recalls, and questionnaires, and with a larger sample.

Finally, the measures of syntactic diversity proposed in this study entail labor-intensive manual analysis, which may impede the practical use of the measures. The development of an

automated analytic tool that can extract incidences of verb-argument constructions would enhance the efficiency and reliability of the analysis. Another possible direction for future research would be to identify early and later developed verb-argument structures and use them as benchmarks for different levels of L2 proficiency. For example, one could investigate the emergence of the construction relative to proficiency level employing the implicational scaling technique. The identified benchmark structures could be used in language pedagogy and assessment.

APPENDICES

Appendix A
English Proficiency Test (C-test)

Fill in <u>one word</u> in each blank. You may write directly on the test. Complete the texts in order (TEXT 1 ➔TEXT 2 ➔TEXT 3).

예) The girl was walking (0) d_____ *the street when she stepped on some ice and fell.*

Answer: down

TEXT 1

Steven loved almost everything about his grandma. There was only one thing he hated. She always knitted sweaters for (1) h_____. Steven understood that she did it to be (2) n_____. However, all the sweaters were very ugly. Steven (3) v_____ her once a week. She had a new (4) s_____ for him each time.

Steven lived in a (5) s_____ apartment. There was no room for him to (6) k_____ all the sweaters. He had to give all of them (7) a_____. "Grandma will never find out," he thought. One (8) d_____, Steven's grandma visited him by surprise. She asked to (9) s_____ his sweaters. "Someone stole all of them!" he (10) s_____. "They were too nice." She (11) m_____ him ten more by the next month.

TEXT 2

Depression is a serious but treatable disorder that affects millions of people, from young to old and from rich to poor. It gets in the way of everyday (12) l_____, causing tremendous pain, hurting not just those suffering (13) f_____ it, but also impacting everyone around them.

If (14) s_____ you love is depressed, you may be (15) e_____ any number of difficult emotions, including helplessness, frustration, (16) a_____, fear, guilt, and sadness. These feelings are all (17) n_____. It's not easy dealing with a friend or (18) f_____ member's depression. And if you don't take care of (19) y_____, it can become overwhelming.

That said, there are (20) s_____ you can take to help your loved one. Start by learning about depression and how to talk (21) a_____ it with your friend or family member. But as you reach out, don't forget to (22) l_____ after your own emotional (23) h_____. Thinking about your own needs is not an (24) a_____ of selfishness—it's a necessity. Your emotional strength will (25) a_____ you to provide the ongoing support your depressed friend or family member needs.

TEXT 3

Nonverbal communication includes facial expressions, gestures, the distance between speakers, eye contact, voice intonations, touch, and many other minor details which can provide speakers with valuable details about each other. For example, (26) s_____ between people can say a lot about the level of intimacy between them: usually, the (27) s_____ the distance between speakers, the more friendly or (28) i_____ they are, and vice versa. Or if a person (29) a_____ eye contact, it might mean that he or she is hiding something, feels (30) u_____ around you, and so on.

Body (31) l_____ has several important functions. For instance, a person's (32) g_____ can repeat the message he or she is (33) m_____ orally; a little child explaining how birds (34) f_____ and waving his or her arms like (35) w_____ is a decent example of this function. Another function, substitution, occurs when (36) v_____ messages can be expressed by nonverbal means (like shrugging). (37) I_____ addition, gestures can be used for accenting, like when (38) r_____ one's index finger when speaking about (39) s_____ important.

At the same time, it is important to remember that sometimes body language may (40) d_____ depending on culture. For example, in some eastern countries, (41) l_____ straight in the eyes of a conversationalist is considered (42) r_____. Men in some Arabic countries may walk around the street (43) h_____ hands, or may kiss each other on the (44) c_____ when greeting, but this is the (45) i_____ of friendship, not romance or intimacy.

Appendix B

Table 24

C-test: Item Facilities and Item Discriminations

Item	IF	IF (Upper)	IF (Lower)	ID
1	0.82	0.95	0.63	0.32
2	0.42	0.78	0.07	0.70
3	0.58	0.91	0.23	0.67
4	0.80	0.96	0.52	0.44
5	0.67	0.89	0.37	0.52
6	0.60	0.85	0.24	0.60
7	0.26	0.65	0.01	0.64
8	0.86	1.00	0.57	0.43
9	0.21	0.37	0.01	0.36
10	0.53	0.63	0.27	0.37
11	0.55	0.94	0.09	0.85
12	0.42	0.82	0.02	0.80
13	0.47	0.78	0.06	0.71
14	0.57	0.95	0.02	0.93
15	0.03	0.13	0.00	0.13
16	0.19	0.47	0.03	0.44
17	0.13	0.24	0.02	0.22
18	0.70	0.97	0.13	0.84
19	0.45	0.76	0.06	0.69
20	0.08	0.24	0.00	0.24
21	0.49	0.88	0.08	0.80
22	0.24	0.64	0.00	0.64
23	0.12	0.29	0.01	0.28
24	0.04	0.15	0.00	0.15
25	0.13	0.33	0.00	0.33
26	0.12	0.39	0.00	0.39

Table 24 (cont'd)

Item	IF	IF (Upper)	IF (Lower)	ID
27	0.27	0.65	0.00	0.65
28	0.35	0.85	0.00	0.85
29	0.17	0.48	0.00	0.48
30	0.33	0.71	0.01	0.70
31	0.62	1.00	0.02	0.98
32	0.54	0.92	0.03	0.89
33	0.07	0.23	0.00	0.23
34	0.36	0.83	0.01	0.82
35	0.24	0.56	0.00	0.56
36	0.38	0.77	0.00	0.77
37	0.57	0.97	0.07	0.90
38	0.10	0.34	0.00	0.34
39	0.46	0.94	0.02	0.92
40	0.18	0.56	0.00	0.56
41	0.35	0.87	0.01	0.86
42	0.39	0.91	0.00	0.91
43	0.17	0.50	0.00	0.50
44	0.27	0.72	0.01	0.71
45	0.06	0.19	0.00	0.19

Appendix C

Language Learning Background Questionnaire (for college students)

The following are questions about your English learning experiences. Read each item carefully, and place a check (√) mark next to the appropriate answer, or fill out with a brief answer.

1. Gender: □Male □Female

2. Year of Study: □Freshman □Sophomore □Junior □Senior

3. Major:_____

5. I studied/am studying English from ___ to ___ year old.

6. Indicate the number that best represents your English proficiency.

	(**1**: Minimal, **2**: Basic, **3**: Good , **4**: Very good, **5**: Excellent)				
• Overall English:	□1	□2	□3	□4	□5
• Reading:	□1	□2	□3	□4	□5
• Writing:	□1	□2	□3	□4	□5
• Speaking:	□1	□2	□3	□4	□5
• Listening:	□1	□2	□3	□4	□5

7. Have you ever lived in an English-speaking country (for example, USA, UK, Canada, Australia, Philippines, Singapore, Hong Kong)? □Yes □No

   If yes:

Age	Country	Length of Residence
Example: years old	*US*	*1year and 2 months*

8. Have you taken a standardized English test (for example, TOEFL, TOEIC, TEPS, IELTS) □Yes □ No
   - Test: _____
   - Approximate date: Year_____ Month _____
   - Score: _____

Appendix D

Language Learning Background Questionnaire (for high school students)

The following are questions about your English learning experiences. Read each item carefully, and place a check (√) mark next to the appropriate answer, or fill out with a brief answer.

1. Gender: □Male □Female

2. Year of Study: □ First year □ Second year □ Third year

3. I studied/am studying English from _____ to _____ year old.

4. Indicate the number that best represents your English proficiency.

	(**1**: Minimal, **2**: Basic, **3**: Good , **4**: Very good, **5**: Excellent)				
• Overall English:	□1	□2	□3	□4	□5
• Reading:	□1	□2	□3	□4	□5
• Writing:	□1	□2	□3	□4	□5
• Speaking:	□1	□2	□3	□4	□5
• Listening:	□1	□2	□3	□4	□5

6. Have you ever lived in an English-speaking country (for example, USA, UK, Canada, Australia, Philippines, Singapore, Hong Kong)? □Yes □No

   If yes:

Age	Country	Length of Residence
Example: 9-10 years old	*US*	*1year and 2 months*

7. Have you taken a standardized English test (for example, TOEFL, TOEIC, TEPS, IELTS) □Yes □ No
   - Test: _____
   - Approximate date: Year_____ Month _____
   - Score: _____

Appendix E

Language Learning Background Questionnaire in Korean (for college students)

언어 학습 배경 설문지 (대학생용)

다음은 본인의 영어 학습 경험에 대한 질문입니다. 각 문항을 잘 읽어보신 후, 그 문항에 알맞은 답에 체크(√) 표시를 하거나 답을 간단하게 서술하여 주십시오.

1. 성별: □남 □여

2. 학년: □1 □2 □3 □4

3. 전공:_____

4. 나는 영어 학습을 _____살부터 _____살 까지 했다/하고 있다.

5. 자신의 영어 능력을 가장 잘 설명하는 숫자를 골라 표기하여 주십시오.

(1: 최소한 , 2: 기초적, 3: 준수한 , 4: 우수한, 5: 탁월한)

- 전반적 영어능력 : □1 □2 □3 □4 □5
- 읽기 능력: □1 □2 □3 □4 □5
- 쓰기 능력: □1 □2 □3 □4 □5
- 말하기 능력: □1 □2 □3 □4 □5
- 듣기 능력: □1 □2 □3 □4 □5

6. 영어권 국가 (예시: 미국, 영국, 호주, 필리핀, 싱가폴, 홍콩) 거주 경험은? □있다 □없다

있다면:

도착 나이	국가	거주기간
Example: 13 세	*미국*	*1 년 2 개월*

7. 토플/토익/텝스/IELTS 등 영어 시험을 보신 적이 있습니까? □예 □ 아니오

- 시험 이름: _____
- 대략적인 시험 날짜: _____년 _____월
- 점수: _____

Appendix F

Language Learning Background Questionnaire in Korean (for high school students)

언어 학습 배경 설문지 (고등학생용)

다음은 본인의 영어 학습 경험에 대한 질문입니다. 각 문항을 잘 읽어보신 후, 그 문항에 알맞은 답에 체크(√) 표시를 하거나 간단하게 서술하여 주십시오.

1. 성별: □남 □여

2. 학년: □1 □2 □3

3. 나는 영어 학습을 _____살부터 _____살 까지 했다/하고 있다.

4. 자신의 영어 능력을 가장 잘 설명하는 숫자를 골라 표기하여 주십시오.

(1: 최소한, 2: 기초적, 3: 준수한 , 4: 우수한, 5: 탁월한)

- 전반적 영어능력 : □1 □2 □3 □4 □5
- 읽기 능력: □1 □2 □3 □4 □5
- 쓰기 능력: □1 □2 □3 □4 □5
- 말하기 능력: □1 □2 □3 □4 □5
- 듣기 능력: □1 □2 □3 □4 □5

5. 영어권 국가 (예시: 미국, 영국, 호주, 필리핀, 싱가폴, 홍콩) 거주 경험은? □있다 □없다

있다면:

도착 나이	국가	거주기간
Example: 13 세	*미국*	*1 년 2 개월*

6. 토플/토익/텝스/IELTS 등 영어 시험을 보신 적이 있습니까? □예 □ 아니오

- 시험 이름: _____
- 대략적인 시험 날짜: _____년 _____월
- 점수: _____

Appendix G

Rater Background Questionnaire

PLEASE FILL OUT THE FOLLOWING BACKGROUND INFORMATION. PLEASE PRINT CLEARLY.

1. Name: a. First name: _____
   b. Last name: _____
2. Age: _____

3. Native language: _____

4. Language you speak at home: _____

5. Are you now or have you ever been an English as a Second or Foreign Language (ESL/EFL) teacher?

   ☐Yes ☐No

   a. If yes, for how long (total)?
   ☐1 year or less ☐2-5 years ☐5-10 years ☐More than 10 years

   b. If yes, what state(s) (US) or country (countries) did you teach in?

   a ._____ How long did you teach there?_____
   b. _____ How long did you teach there?_____
   c. _____ How long did you teach there?_____

6. Do you have previous experience rating ESL/EFL compositions?
   ☐Yes ☐No

   a. If yes, could you briefly describe your experience?
   _____
   _____
   _____

7. How do you describe your abilities to evaluate ESL/EFL compositions?
   ☐ Novice ☐ Competent ☐ Excellent

8. What languages, other than English, do you speak or have you studied or are currently studying? Please report and answer questions for each language other than English that you speak or have studied or are currently studying.

LANGUAGE A.	HOW DID YOU LEARN THE LANGUAGE? (Please describe.)	From what age to what age did you learn the language? ___ to ___	HOW WELL DO YOU SPEAK THE LANGUAGE? (Please circle one) poor / fair / good / advanced/ fluent / native-like Comments:
LANGUAGE B.	HOW DID YOU LEARN THE LANGUAGE? (Please describe.)	From what age to what age did you learn the language? ___ to ___	HOW WELL DO YOU SPEAK THE LANGUAGE? (Please circle one) poor / fair / good / advanced/ fluent / native-like Comments:

9. Have you lived in or traveled to a place where people speak the languages you speak or have studied or are currently studying (the ones listed in #9)?

☐Yes ☐No

If ***yes***, please report and answer questions for each place you have lived or visited and where the language(s) (#10) were spoken.

Where did you travel or live? a. _____	For how long were you there?	How old were you when you were there?	What was the purpose of your visit or stay?
Where did you travel or live? a. _____	For how long were you there?	How old were you when you were there?	What was the purpose of your visit or stay?

Appendix H

Table 25

Rating rubric

	Content		Organization		Vocabulary		Language Use	Score /2	Mechanics
20 16	Thorough and logical development of thesis Substantive and detailed No irrelevant information Interesting A substantial number of words for amount of time given	20 16	Excellent overall organization Clear thesis statement Substantive introduction and conclusion Excellent use of transition word Excellent connections between paragraphs Unity within every paragraph	20 16	Very sophisticated vocabulary Excellent choice of words with no errors Excellent range of vocabulary Idiomatic and near native-like vocabulary Academic register	20 16	No major errors in word order or complex structures No errors that interfere with comprehension Only occasional errors in morphology Frequent use of complex sentences Excellent sentence variety	20 16	Appropriate layout with indented paragraphs No spelling errors No punctuation errors
15 11	Good and logical development of thesis Fairly substantive and detailed Almost no irrelevant information Somewhat interesting An adequate number of words for the amount of time given	15 11	Good overall organization Clear thesis statement Good introduction and conclusion Good use of transition words- Good connections between paragraphs Unity within most paragraphs	15 11	Somewhat sophisticated vocabulary Attempts, even if not completely successful, at sophisticated vocabulary Good choice of words with some errors that don't obscure meaning Adequate range of vocabulary but some repetition Approaching academic register	15 11	Occasional errors in awkward order or complex structures Almost no errors that interfere with comprehension Attempts, even if not completely successful, at a variety of complex structures Some errors in morphology Frequent use of complex sentences Good sentence variety	15 11	Appropriate layout with indented paragraphs No more than a few spelling errors in less frequent vocabulary No more than a few punctuation errors
10 6	Some development of thesis Not much substance or detail Some irrelevant information Somewhat uninteresting Limited number of words for the amount of time given	10 6	Some general coherent organization Minimal thesis statement or main idea Minimal introduction and conclusion Occasional use of transitions words Some disjointed connections between paragraphs Some paragraphs may lack unity	10 6	Unsophisticated vocabulary Limited word choice with some errors obscuring meaning Repetitive choice of words No resemblance to academic register	10 6	Errors in word order or complex structures Some errors that interfere with comprehension Frequent errors in morphology Minimal use of complex sentences Little sentence variety	10 6	Appropriate layout with most paragraphs indented Some spelling errors in less frequent and more frequent vocabulary Several punctuation errors
5 0	No development of thesis No substance or details Substantial amount of irrelevant information Completely uninteresting Very few words for the amount of time given	5 0	No coherent organization No thesis statement or main idea No introduction and conclusion No use of transition words Disjointed connections between paragraphs Paragraphs lack unity	5 0	Very simple vocabulary Severe errors in word choice that often obscure meaning No variety in word choice No resemblance to academic register	5 0	Serious errors in word order or complex structures Frequent errors that interfere with comprehension Many error in morphology Almost no attempt at complex sentences No sentence variety	5 0	No attempt to arrange essay into paragraphs Several spelling errors even in frequent vocabulary Many punctuation errors

REFERENCES

REFERENCES

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.), *Automatic treatment and analysis of learner corpus data* (pp. 249-264). Amsterdam/ Philadelphia: John Benjamins.

Alishahi, A., & Stevenson, S. (2008). A computational model of early argument structure acquisition. *Cognitive science, 32*(5), 789-834.

Anthony, L. (2014). AntConc (Version 3.4.3) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/

Anthony, L. (2014). TagAnt (Version 1.1.2) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/

Asención-Delaney, Y., & Collentine, J. (2011). A multidimensional analysis of a written L2 Spanish corpus. *Applied Linguistics, 32*(3), 299-322.

Babaii, E., & Ansary, H. (2001). The C-test: a valid operationalization of reduced redundancy principle?. *System, 29*(2), 209-219.

Bachman, L. F. (1985). Performance on cloze tests with fixed-ratio and rational deletions. *TESOL Quarterly, 19*(3), 535-556.

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

Baralt, M. (2012). Coding qualitative data. In A. Mackey & S. Gass (Eds.), *Research methods in second language acquisition* (pp. 222-244). Malden, MA: Wiley-Blackwell.

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

Barkaoui, K. (2010). Variability in ESL essay rating processes: The role of the rating scale and rater experience. *Language Assessment Quarterly, 7*(1), 54-74.

Becini, G. M. L. & Goldberg, A. (2000). The contribution of argument structure constructions to sentence meaning. *Journal of Memory and Language 43*, 640-651.

Benevento, C., & Storch, N. (2011). Investigating writing development in secondary school learners of French. *Assessing Writing, 16*(2), 97-110.

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.

Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E., & Quirk, R. (1999). *Longman grammar of spoken and written English* (Vol. 2). MIT Press.

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

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. 21-46). John Benjamins Publishing.

Bulté, B., & Housen, A. (2014). Conceptualizing and measuring short-term changes in L2 writing complexity. *Journal of Second Language Writing*, *26*, 42-65.

Byrnes, H., Maxim, H. H., & Norris, J. M. (2010). Introduction. *The Modern Language Journal*, *94*(s1), 1-202.

Canale, M. & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics 1*(1), 9-47.

Carroll, J. B. (1964). *Language and Thought*. Englewood Cliffs, NJ: Prentice-Hall .

Casanave, C. P. (1994). Language development in students' journals. *Journal of second language writing*, *3*(3), 179-201.

Chapelle, C. A., & Duff, P. A. (2003). Some guidelines for conducting quantitative and qualitative research in TESOL. *TESOL quarterly*, *37*(1), 157-178.

Cohen, J. (1992). A power primer. *Psychological bulletin*, *112*(1), 155.

Connor-Linton, J., & Polio, C. (2014). Comparing perspectives on L2 writing: Multiple analyses of a common corpus. *Journal of Second Language Writing*, *26*, 1-9.

Cooper, T. C. (1976). Measuring written syntactic patterns of second language learners of German. *The Journal of Educational Research*, *69*(5), 176-183.

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.Carr 2011

Crossley, S. A., & McNamara, D. S. (2011). Shared features of L2 writing: Intergroup homogeneity and text classification. *Journal of Second Language Writing, 20*, 271-285.

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.

Cumming, A., Kantor, R., & Powers, D. E. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *The Modern Language Journal*, *86*(1), 67-96.

Danzak, R. L. (2011). The integration of lexical, syntactic, and discourse features in bilingual adolescents' writing: An exploratory approach. *Language, Speech, and Hearing Services in Schools*, *42*(4), 491-505.

Ellis, N. & Ferreira-Junior, F. (2009a). Constructions and their acquisition. *Annual Review of Cognitive Linguistics, 7*, 187-220.

Ellis, N. & Ferreira-Junior, F. (2009b). Construction learning as a function of frequency, frequency distribution and function, *Modern Language Journal, 93*(3), 370-385.

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

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.

Foster, P., & Skehan, P. (1996). The influence of planning and task type on second language performance. *Studies in Second language acquisition*, *18*(3), 299-323.

Frear, M. W., & Bitchener, J. (2015). The effects of cognitive task complexity on writing complexity. *Journal of Second Language Writing*, *30*, 45-57.

Friedman, D. A. (2012). How to collect and analyze qualitative data. In A. Mackey & S. Gass (Eds.), *Research methods in second language acquisition* (pp. 188-200). Malden, MA: Wiley-Blackwell.

Goldberg, A. (1995). *Constructions: A construction grammar approach to argument structure*. Chicago, IL: University of Chicago Press.

Goldberg, A., & Suttle, L. (2010). Construction grammar. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(4), 468-477.

Grant, L., & Ginther, A. (2000). Using computer-tagged linguistic features to describe L2 writing differences. *Journal of second language writing*, *9*(2), 123-145.

Guo, L., Crossley, S. A., & McNamara, D. S. (2013). Predicting human judgments of essay quality in both integrated and independent second language writing samples: A comparison study. *Assessing Writing, 18,* 218-238.

Gyllstad, H., Granfeldt, J., Bernardini, P., & Kallkvist, M. (2014). Linguistic correlates to communicative proficiency levels of the CEFR. *EUROSLA Yearbook, 14*, 1-30.

Hinkel, E. (2003). Simplicity without elegance: Features of sentences in L1 and L2 academic texts. *TESOL Quarterly*, *37*(2), 275-301.

Homburg, T. J. (1984). Holistic evaluation of ESL compositions: Can it be validated objectively?. *TESOL quarterly*, *18*(1), 87-107.

Housen, A., & Kuiken, F. (2009). Complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics*, *30*(4), 461-473.

Housen, A., Kuiken, F., & Vedder, I. (Eds.). (2012). *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (Vol. 32). John Benjamins Publishing.

Hunt, K. W. (1965). Grammatical structures written at three grade levels. NCTE Research Report No. 3.

Hunt. K. W. (1970). Recent measures in syntactic: development. In M. Lester (Ed.), *Readings in applied transformational grammar* (pp. 187-200). New York: Holt, Rinehert.

Ishikawa, S. (1995). Objective measurement of low-proficiency EFL narrative writing. *Journal of Second Language Writing*, *4*(1), 51-69.

Ishikawa, T. 2007. The effect of manipulating task complexity along the +/- here-and-now dimension on L2 written narrative discourse. In M. P. Garcia Mayo (Ed.), *Investigating tasks in formal language learning.* Multilingual Matters.

Jacobs, H., Zinkgraf, S., Wormuth, D., Hartfiel, V., & Hughey, J. (1981). *Testing ESL Composition: A practical approach.* Rowley, MA: Newbury House.

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.

Kim, S. H. (2014). *Metacognitive knowledge in second language writing* (Unpublished doctoral dissertation). Michigan State University.

Klecka, W. R. (1980). *Discriminant analysis*. Beverly Hills, CA: Sage.

Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing*, *26*(2), 275-304.

Kormos, J., & Trebits, A. (2012). The role of task complexity, modality, and aptitude in narrative task performance. *Language Learning*, *62*(2), 439-472.

Kuiken, F., & Vedder, I. (2007). Task complexity and measures of linguistic performance in L2 writing. *IRAL-International Review of Applied Linguistics in Language Teaching*, *45*(3), 261-284.

Kuiken, F., & Vedder, I. (2014). Rating written performance: What do raters do and why?. *Language Testing*, *31*(3), 329-348. Landers 2015

Larsen-Freeman, D. (2006). The emergence of complexity, fluency, and accuracy in the oral and written production of five Chinese learners of English. *Applied Linguistics, 27*(4), 590-519.

Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language Learning 40*, 387-417.

Llanes, A., & Munoz, C. (2013). Age effects in a study abroad context: Children and adults studying abroad and at home. *Language Learning*, *63*(1), 63-90.

Long, S. H., Fey, M. E., & Channell, R. W. (2008). Computerized Profiling (CP) (Version 9.2. 7, MS-DOS)[computer program]. *Cleveland, OH: Department of Communication Sciences, Case Western Reserve University*.

Lu, X. (2009). Automatic measurement of syntactic complexity in child language acquisition. *International Journal of Corpus Linguistics*, *14*(1), 3-28.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics, 15*(4), 474-496

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

Lu, X., & 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.

Lumley, T. (2002). Assessment criteria in a large-scale writing test: what do they really mean to the raters?. *Language Testing*, *19*(3), 246-276.

Lumley, T. (2005). *Assessing second language writing: The rater's perspective*. Frankfurt: Peter Lang.

Mancilla, R. L., Polat, N., & Akcay, A. O. (2015). An investigation of native and nonnative English speakers' levels of written syntactic complexity in asynchronous online discussions. *Applied Linguistics*, http://dx.doi.org/10.1093/applin/amv012

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.

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

Monroe, J. H. (1975). Measuring and enhancing syntactic fluency in French. *The French Review*, *48*(6), 1023-1031.

Nelson, L. R. (2000). *Item analysis for tests and surveys using Lertap 5*. Perth, Western Australia: Curtin University of Technology.

Norrby, C. & Hakansson, G. (2007). The interaction of complexity and grammatical processability: The case of Swedish as a foreign language. *International Review of Applied Linguistics, 45*, 45-68.

Norris, J. M. (2015). Discriminant analysis. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 305-328). New York/ London: Routledge.

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.

Norris, J., & Ortega, L. (2003). Defining and measuring SLA. In C. Doughty, & M. Long (Eds.), *The handbook of second language acquisition* (pp. 716-761). John Wiley & Sons.

Oller, J. W., Jr . (1972). Scoring methods and difficulty levels for cloze tests of proficiency in English as a second language. *Modern Language Journal, 56*, 151–157.

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.

Ortega, L. (2012). Interlanguage complexity: A construct in search of theoretical renewal. In B. Kortmann & B. Szmrecsanyi (Eds.), *Linguistic complexity: Second language acquisition, indigenization, contact* (pp. 127–155). Berlin: De Gruyter.

Palloti, G. (2015). A simple view of linguistic complexity. *Second Language Research, 31*(1), 117-134.

Pallotti, G. (2009). CAF: Defining, refining and differentiating constructs. *Applied Linguistics, 30*(4), 590-601.

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). Mahwah, NJ: Lawrence Erlbaum.

Polio, C. (2013). *Revising a writing rubric based on raters' comments: Does it result in a more reliable and valid assessment?*.Midwest Association of Language Testers, Michigan State University.

Purpura, J. (2004). *Assessing grammar*. Cambridge: Cambridge University Press.

Quirk, R., Greenbaum, S., Leech, G., Svartvik, J., & Crystal, D. (1985). *A comprehensive grammar of the English language* (Vol. 397). London: Longman.

Raatz, U. & Klein-Braley, C. (1981). The C-test: A modification of the cloze procedure. In T. Culhane, C. Klein-Braley, & D. K. Stevenson (Eds.), *Practice and problems in language testing* (pp. 113-138). Colchester: Department of Language and Linguistics, University of Essex.

Ramos, S. D. S., & Rickard Liow, S. J. (2013). Discriminant function analysis. *The Encyclopedia of Applied Linguistics*.

Révész, A. (2008). Task complexity, focus on form-meaning connections, and individual differences: A classroom-based study. Paper presented at the *International Association of Applied Linguistics*, Essen, Germany.

Rimmer, W. (2006). Measuring grammatical complexity: the Gordian knot. *Language Testing, 23*(4), 497-519

Rimmer, W. (2008). Putting grammatical complexity in context. *Literacy, 42*(1), 29-35.

Serrano, R., Llanes, A., & Tragant, E. (2011). Analyzing the effect of context of second language learning: Domestic intensive and semi-intensive courses vs. study abroad in Europe. *System, 29*, 133-143.

Serrano, R., Tragant, E., & Llanes, A. (2012). A longitudinal analysis of the effects of one year abroad. *The Canadian Modern Language Review, 68*(2), 183-163.

Shang, H.-F. (2007). An exploratory study of e-mail application on FL writing performance. *Computer Assisted Language Learning, 20*(1), 79-96.

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

Spoelman, M., & Verspoor, M. (2010). Dynamic patterns in development of accuracy and complexity: A longitudinal case study in the acquisition of Finnish. *Applied Linguistics, 31*(4), 532-553.

Stockwell, G. (2005). Syntactical and lexical development in NNS-NNS asynchronous CMC. *The JALT CALL Journal, 1*(3), 33-49.

Stockwell, G., & Harrington, M. (2003). The incidental development of L2 proficiency in NS-NNS email interactions. *CALICO journal*, *20*(2) 337-359.

Storch, N. (2009). The impact of studying in s second language (L2) medium university on the devlopment of L2 writing. *Journal of Second Language Writing, 18,* 103-118.

Storch, N., & Tapper, J. (2009). The impact of an EAP course on postgraduate writing. *Journal of English for Academic Purposes, 8*(3), 207-223.

Storch, N., & Wigglesworth, G. (2007). Writing tasks: The effects of collaboration. In M. P. Garcia Mayo (Ed.), *Investigating tasks in formal language learning* (pp. 157-177). Clevedon: Multilingual Matters.

Taguchi, N., Crwoford, 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.

Tremblay, A. (2011). Proficiency assessment standards in second language acquisition research. *Studies in Second Language Acquisition*, *33*(3), 339-372.

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.

Vyatkina, N. (2012). The development of second language writing complexity in groups and individuals: A longitudinal learner corpus study. *The Modern Language Journal, 96*(4), 576-598.

Vyatkina, N. (2013). Specific syntactic complexity: Developmental profiling of individuals based on an annotated learner corpus. *The Modern Language Journal, 97*(s1), 11-30.

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.

Winke, P., & Gass, S. (2013). The influence of second language experience and accent familiarity on oral proficiency rating: A qualitative investigation. *TESOL Quarterly*, *47*(4), 762-789.

Winke, P., & Lim, H. (2015). ESL essay raters' cognitive processes in applying the Jacobs et al. rubric: An eye-movement study. *Assessing Writing*, *25,* 38-54.

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

Yoon, H. J., & Polio, C. (2016). The linguistic development of students of English as a second Language in two written genres. *TESOL Quarterly*. doi:10.1002/tesq.296

Zandi, H. (2014). Investigating the relationship among complexity, range, and strength of grammatical knowledge of EFL students. *Applied Research on English Language, 3*(2), 85-100.