ASSESSING GRAMMATICAL FEATURES ACROSS SCORE LEVELS IN SECOND LANGUAGE WRITING: A CORPUS-BASED ANALYSIS

By

Susie Kim

A DISSERTATION

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

Second Language Studies—Doctor of Philosophy

2019

ABSTRACT

ASSESSING GRAMMATICAL FEATURES ACROSS SCORE LEVELS
IN SECOND LANGUAGE WRITING: A CORPUS-BASED ANALYSIS

By

Susie Kim

Recent research in the areas of second language testing and learner corpus research has provided increased insight into linguistic features of various score levels and into the meaning of a test score (Cushing, 2017; Knoch & Chapelle, 2018). However, language testing researchers have asserted the need to select linguistic features that are relevant to the test construct for test validation purposes (Egbert, 2017; Xi, 2017). In addition, the Common European Framework of Reference (CEFR) has been widely adopted in testing contexts and provides level descriptions for linguistic abilities, but empirical validation of its use in various testing contexts is critical (Wisniewski, 2017, 2018). Addressing these two limitations, I drew upon learner-produced written English from a large-scale English exam, the *Certificate of English Language Competency* (CEFR B2-level certification). The aim of the study was to (a) investigate specific grammatical features and overall linguistic accuracy of second language English texts to reveal patterns of language use at different score levels, and (b) examine how well rating rubric descriptors reflect characteristics of examinee texts and differentiate between score levels to find evidence for test validity.

In order to provide concrete, context-relevant grammatical features for investigation, I selected 14 grammatical features from the *English Profile* studies (English Profile, 2015; Hawkins & Filipović, 2012), which have also been documented in L2 writing research. Data included 560 texts written on three different topics and ranging across five levels of performance. I extracted the occurrences of 14 grammatical features from the corpus using

Natural Language Processing tools and analyzed occurrences of these features attested in the corpus. Additionally, a subset of the texts was manually coded for error to examine overall accuracy of each texts.

Consistent with the findings in existing literature, I found significant differences in the frequencies of certain clausal features across lower score levels. Both the frequencies of the 14 grammatical features and the overall number of different types of these features used in each text were moderately useful in predicting the grammar subscore. I identified co-occurring patterns of the target grammatical features by performing a principal components analysis. The results showed that grammar structures that are of similar types (e.g., finite, non-finite) and functions (e.g., complement, noun modifier) tended to occur together and exhibited (cross-sectional) developmental patterns. For a subset of data coded for errors, the error-free clause ratio was calculated, which significantly distinguished between each pair of adjacent levels.

This study's findings highlight the need for empirical investigation of how learner language has been described by experts in proficiency descriptors (e.g., Council of Europe, 2001, 2018) and how reliably the constructs of rubric descriptors attest in test performance data. I suggest that writing assessment materials can benefit from reference to the tangible characteristics of L2 development found in writing development research (e.g., phrasal complexity, morphological accuracy, and association strength between a construction and its lexis).

For Mom
Your smile and warmth keep me strong.

ACKNOWLEDGEMENTS

I would like to express my gratitude to all faculty, students, and alumni of the Second Language Studies program. I am blessed to have joined such a tight-knit community full of positive energy, support, intellect, and diverse expertise.

I am especially grateful to my dissertation co-chairs, Dr. Charlene Polio and Dr. Daniel Reed, for their guidance and insight throughout the process of completing this dissertation. Their warm encouragement and reassurance helped me make progress. I am indebted to Dr. Paula Winke for her mentorship in my successfully navigating the program. I also thank Dr. Sandra Deshors for her valuable perspective in improving my dissertation.

I acknowledge the English Language Center Testing Office and its staff for granting the access to MSU Exams data for this dissertation and the College of Arts and Letters for financial aids. I am thankful to Ann Letson and Amy Cheadle for turning the frustrating task of data analysis into fun discussions and discoveries. I would also like to thank the Korean program at MSU and Dr. Ok-Sook Park for offering me the priceless experience in teaching, which served as a delightful release from toiling over this dissertation.

Special thanks go out to Ji-Hyun Park, Dustin Crowther, Magda Tigchelaar, Zack Miller, Dan Isbell, Jungmin Lim, Wendy Li, Matt Kessler, and Unhee Ju for their friendship, advice, and humor. I enjoyed every day in this program because of these friends who were always there to believe in me and cheer me up.

The more feats I achieve in my life, the more I grow grateful to my family for their infinite love and support. I am forever thankful to Dad, Sangho, and both of my grandmas. I am proud to follow in my parents' footsteps to be a scholar and a teacher.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Corpus linguistic data has increasingly been used in language testing in recent years, as it serves the interests of language testers in test validation by providing opportunities to expound test score and usefulness with linguistic evidence. Specifically, a body of corpus-informed research revealing linguistic features of score levels has provided a means to make inferences regarding test score and linguistic ability (e.g., Cumming, Kantor, Baba, Eouanzoui, Erdosy, & James, 2005; Knoch, Macqueen, & O'Hagan, 2014; Kyle & Crossley, 2017; LaFlair & Staples, 2017). Adopting learner corpora and corpus-based analysis provides a number of benefits for assessment purposes: instructing language testers on the linguistic features that are characteristic of learners' proficiency levels (Park, 2014), creating authentic assessment tasks and improving assessment criteria and scales (Taylor & Barker, 2008), enhancing the understanding of test construct and validity by analyzing real language across levels (Barker, Salamoura, & Saville, 2015), and advancing automated scoring systems (Cushing, 2017), to name a few. Consequently, scholars in learner corpus research and language testing alike have actively been underscoring the benefits and methods of research that employs learner corpora for second language assessment purposes (see Barker et al., 2015; Callies et al., 2014), most recently with Cushing (2017) highlighting the numerous ways corpus-based analysis aids test validity. Even though this transdisciplinary research has been offering novel approaches to finding evidence for test validity, language testing researchers have pointed out that the selection of linguistic features for test validation needs to be motivated by the test construct to be meaningful (Egbert, 2017; Xi, 2017). At the same time, widely used proficiency scales such as the Common European Framework of Reference (CEFR) still call for empirical validation in various testing contexts (Wisniewski, 2017, 2018).

Empirical research into learner-produced English language in testing contexts in relation to the CEFR scale has been largely limited in two regards. Firstly, most research has been conducted only using the *Cambridge Learner Corpus*. Although it is a large corpus that includes learners from many different language backgrounds, the literature could benefit from exploration of both data collected in a context other than the Cambridge English exams and data produced by first language groups that are under-represented in this corpus. Secondly, previous corpus-based investigations into the CEFR levels have mostly been descriptive and have fallen short of making direct connections or implications back to the language test itself. In other words, studies exploring the relationship between learner language and CEFR-related assessment materials (e.g., rating rubrics) or procedures (e.g., scoring) have been insufficient in number and scope despite the CEFR scale's popular use as an assessment tool.

From the perspective of language testing research, inquiries based on methods from applied linguistics (e.g., corpus linguistics) can provide strong support for test validation (Chapelle, 2018; Cushing, 2017). Bachman (1990) noted that, in measurement, the definition of "ability" is complicated as it poses a circularity problem; that is, the ability that a test intends to measure is defined by an individual's performance on the test, but how the individual performs is contingent on his or her ability. Therefore, the specifications of proficiency scale and proficiency descriptions merit empirical examination. In this study, I expand on the research into learner language characteristics and test score interpretation by (a) examining learner-produced texts from a large-scale English exam aligned with the CEFR and (b) evaluating how well the rating rubric and rater-assigned scores reflect learner language. Specifically, I provide a focused investigation into the grammatical structures that researchers have previously found to be reflective of the CEFR proficiency levels (e.g., O'Keeffe & Mark, 2017), thereby creating direct

links among the exam, rating scale, and learner language. Furthermore, I seek empirical evidence for the scoring rubric's relevance to examinee-produced texts and its ability to differentiate between different score levels. The results of this study will contribute to enlarging our understanding of what learners can do with grammar at different levels of proficiency and to strengthening the existing descriptions of the CEFR levels, which will serve as a valuable resource for examinees, teachers, raters and test developers. To achieve these goals, I examine the learner English produced and rated for the *Certification for English Competency* (CELC), a test aligned to CEFR B2 level. I focus on the grammar use by the examinees, in light of the burgeoning interest in corpus-informed language test validation and the continued search for linguistic features that describe and characterize the CEFR proficiency levels. In the following section, I review the argument-based validity framework and studies that have investigated learner-produced texts in relation to score levels or proficiency levels. I also summarize the research findings on the relationship between grammatical features of learner English and score levels.

CHAPTER 1: BACKGROUND

In this section, I survey previous scholarship pertinent to language test validation through examining linguistic characteristics, focusing on text analysis of grammatical features. Related aspects of this study—(a) test validation, (b) relationship between score and linguistic features, and (c) studies on linguistic features of the Common European Framework of Reference levels—are introduced and discussed to situate the current study.

1.1 Test Score Interpretation and Use

1.1.1 The concept of argument-based language test validation

The traditional definition of validity—that it is "the degree to which a test measures what

it is supposed to measure" (Sireci, 2009)—has limited test validity research to simply recognizing a good test and a bad test. In more recent years, scholars have begun to view test validity not as pertaining to the test itself, but to the interpretations of the test score. Messick (1989) seminally defined validity as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores" (p. 13). Following this idea, Kane (1992, 2011, 2013) proposed a praxis framework needed to validate the interpretation of test score and use. In this proposed framework, validation is viewed as the process of building and supporting an argument made about the test score and its use. This argument-based approach has been adopted in the field of language testing, most notably in Chapelle, Enright, and Jamieson's (2008) study on the Test of English as a Foreign Language (TOEFL). The TOEFL is designed to measure academic English proficiency at the university level; hence, the validity argument for this test will center around the interpretation of TOEFL score as the degree of "test takers' language readiness for academic study at English-medium universities" (Chapelle et al., 2008, p. 320). In validating this test, the researchers placed emphasis on developing practical steps to examine the level of support for the interpretation of test score (e.g., language readiness in academic settings) and its intended use (e.g., admission for English-medium universities).

As mentioned, Kane (1992, 2001, 2006) outlined an influential argument-based approach that requires laying out claims about the meaning of test scores and evaluating the claims with supporting evidence. The claims regarding score interpretation are multifaceted, in that there are various components and steps involved in arriving at a decision (e.g., acceptance or non-acceptance decided by a TOEFL score) based on an observation (e.g., a student's language performance on TOEFL tasks). Figure 1.1 illustrates such components (listed in the top row) and

inferences to be made (listed in the bottom row) in a progressive manner, which are pertinent to the interpretation and use of the TOEFL. Starting from the left-most component, the target language use domain for the TOEFL would be in a university classroom where English is spoken. Therefore, TOEFL speaking tasks aim to approximate in-class speaking activities so that a student's speaking ability in this setting can be measured through the student's performance on the test. To make a sound interpretation of the test score, what needs to be examined is whether the observations of test performance reflect student's abilities in situations representative of the target domain (i.e., speaking in English in a university class). In other words, we want to be able to make an inference about the student's speaking ability in the target domain by observing his or her performance on the test. This is the *domain description* inference (see the left-most box in the bottom row of Figure 1.1), which links the target domain and observation of test performance.

Test performance is then translated into an observed score during the rating process. At this stage, an examination of the evaluation process of the test is required in order to infer that the observed score (i.e., the TOEFL speaking score) reflects target language abilities (i.e., English speaking ability).

Then, to help determine how generalizable the observed score is, the next step must be an investigation into whether this score would hold consistent in parallel versions of the test and/or in a retaking of the test (i.e., where a *generalization* inference can be justified).

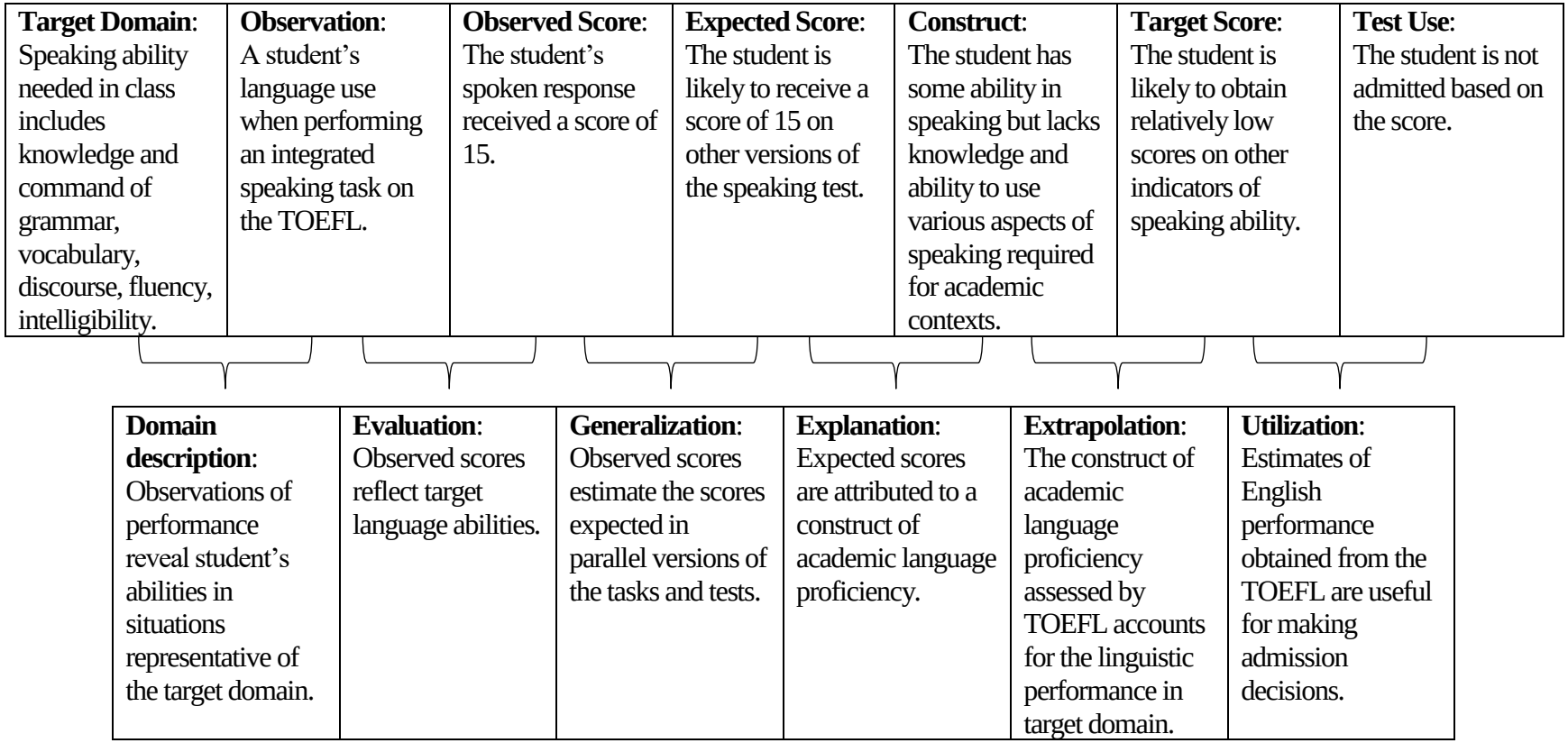

Figure 1.1. *Illustration of inferences in the TOEFL (a summary of Chapelle et al., 2008, pp. 15, 19-20).*

The next inference needed to correctly interpret the score is what the score tells us about the student's ability with regard to the target construct. That is, we want to infer the level of academic speaking ability by looking at the score of 15 (i.e., making an *explanation* inference), and to do this, the score must explain the test construct. In other words, the score should give an idea of what this student can and cannot do and how this student would be different from someone who received a higher score or a lower score.

As the TOEFL intends to predict the level of performance in an academic setting, how well the academic language proficiency assessed by the TOEFL accounts for the linguistic performance in English-medium university settings also needs to be evaluated (i.e., making an *extrapolation* inference).

Lastly, when the TOEFL score is used for decision making in the university admission process, the decision makers infer the student's linguistic performance at the university from the TOEFL score. The usefulness of the TOEFL score as admission requirement/criterion thus requires interpretation and examination (i.e., assessing a *utilization* inference).

In summary, the inferences link or bridge the chain of interpretations regarding test score and use. This practice of outlining inferences that are criterial to score interpretation serves as a solid framework for validity research, where each inference and its main claim are investigated to theoretically or empirically support test validity.

1.1.2 Rating linguistic performance in a validity argument

As discussed, Chapelle et al. (2008) undertook the concepts of an argument-based approach and applied them to an actual language test validation process, a validity argument for the TOEFL. Another informative study that situates test validation under an argument-based approach is Knoch and Chapelle's (2018) analysis of rating processes as part of test validation.

As in Chapelle et al.'s (2008) study, Knoch and Chapelle (2018) specify six inferences for a validity argument regarding test score interpretation and rating of examinee performance. Drawing on scholarly journal articles and testing reports (e.g., on TOEFL, IELTS), they conducted a systematic research synthesis to identify the best practices and research methods for argument-based test validity research. Table 1.1 is an example of a detailed outline of a validity argument specific to the explanation inference.

Explanation inference refers to the claim that expected scores are attributed to a construct of language proficiency. To examine how well this claim holds, a more detailed account of the "underlying premise, rule, or principle," called a *warrant*, needs to be specified (Chapelle et al., 2010, p. 6). Because rating criteria and scale mediate performance and score—the two key components that explanation inference links—the alignment among the two components and rating criteria and scale need to be explicated. To this end, Knoch and Chapelle identified two warrants that would "warrant" the claim regarding the explanation inference: (a) The rating criteria need to reflect the construct of language proficiency (Warrant A in Table 1.1) and (b) the scale needs to distinguish examinees' performance (Warrant B). However, these warrants alone are still not specific enough to drive validation research. Researchers therefore also need to identify *assumptions*, which often serve as research questions that operationalize the warrants, ultimately to provide support for the inference being claimed.

Table 1.1

Explanation Inference and Its Warrants, Assumptions, and Sources for Backing (Knoch & Chapelle, 2018, p. 489)

Explanation inference: Expected scores are attributed to a construct of language proficiency.		
Warrants	Assumptions	Sources for backing
A. The rating criteria are based on a clearly defined construct.	1. The rating scale is based on a defensible theoretical or pedagogical model of proficiency and/or development.	Expert review of scale content; review of test development documentation
	2. Rating scale criteria and descriptors cover the construct (i.e., no construct irrelevance or underrepresentation).	Expert review of scale content and test development documentation; interviews with raters or other experts
	3. Raters' cognitive processes are consistent with the theoretical model of proficiency and/or development.	Raters' verbal protocols show that raters draw on the key aspects underlying the theoretical model of proficiency and/or development
B. The descriptors in the rating scale, which is reflective of the theoretical construct, are identifiable in the candidates' discourse in the response.	4. Test takers' discourse is reflective of the descriptions of performance in the rating rubric.	Discourse analysis of candidate discourse
	5. Relevant features of candidates' discourse differentiate between score levels.	Quantitative analysis of candidate discourse features across score levels

The first warrant for the explanation inference in Knoch and Chapelle (2018) is often evaluated by experts by examining the scale content (e.g., a model of a scale such as the CEFR proficiency levels, which identifies three categories of basic, independent, and proficient user; and six levels from A1 to C2) and test documents (e.g., test specifications) in order to ensure that the rating criteria reflect constructs that the test intends to measure. The second warrant entails an assumption that the candidate's performance reflects the descriptors of the rating rubric. For example, the descriptor from the CEFR written assessment rubric (Council of Europe, 2018) for the B2 level states that a B2-level language user "has a sufficient range of language to be able to

give clear description, express viewpoints on most general topics, using some complex sentence forms" (p. 157). If a linguistic analysis on texts produced by B2-level candidates finds that the attributes illustrated in the descriptors (e.g., complex sentence forms) exist in the actual learner texts that passed the B2 level, it would support the validity of the test. On the other hand, if the analysis finds that texts produced by B2-level learners do not contain complex sentence forms, as suggested by the rubric descriptor, then the meaning of the B2-level linguistic ability becomes obscure. In the next section, I review previous studies that have addressed the relationship between linguistic features of examinees' performance and rating scale.

1.1.3 Previous studies on linguistic features in test validation studies

The developers of major tests of English language proficiency (i.e., TOEFL and IELTS) have commissioned studies to investigate the relationship between linguistic features and rating scores for validation purposes. Table 1.2 shows summaries of such studies including the context and design (i.e., variables included in the study), linguistic features investigated, and significant findings pertaining to the score levels. These studies typically also examined the effect of task types, forms, or first languages (L1s) along with the score levels. For example, Cumming et al. (2005) compared the compositions for independent and integrated writing tasks for the TOEFL. They coded 23 linguistic features in the categories of lexical complexity, syntactic complexity, grammatical accuracy, argument structure, orientations to evidence, and verbatim uses of source text. The researchers also compared three groups of examinees who scored 3, 4, and 5 on the writing section of the test (on a scale from 1 to 5), and observed a tendency, descriptively speaking, that more advanced examinees performed better on most of the measures. The only measure that statistically significantly distinguished between the three proficiency levels, however, was grammatical accuracy. Though these results may lend some support to the validity

of the scoring rubrics and levels, the number of examinees was too small (36 examinees producing essays on six tasks) to make generalizations.

Knoch et al. (2014) conducted a similar study with a larger number of TOEFL compositions and with more diverse measures. One of their aims was to discover features of written responses that were characteristic of different score levels and use this evidence to validate the rating scale. They analyzed 24 features related to accuracy, fluency, syntactic complexity, lexical complexity, coherence, cohesion, content, orientation to source evidence, and metadiscourse. Data included 480 examinees' compositions spread across the five score levels. The researchers evaluated how well each feature distinguished between score levels in relation to the rubric descriptors. They reported that the number of ideas in the essay, density of patterns (i.e., native-like formulaic structures), accuracy (i.e., error-free T-units and clauses), and number of words strongly distinguished the score levels.

Biber and Gray's (2013) study of the TOEFL iBT was more comprehensive in the sense that they examined both speaking and writing performances and incorporated an extensive number of features. This study was also different in that the linguistic features under investigation were specific lexico-grammatical items and not the traditional complexity measures. By accounting for the interaction between different variables (i.e., mode, task, and score level), the researchers successfully made the point that frequently-occurring linguistic features differed depending on the mode (i.e., speaking versus writing). However, they found little relationship between grammatical features and score level.

Table 1.2

Studies Investigating Linguistic Features for Test Validation

Study	Context and design	Linguistic features	Significant findings for score levels
Cumming et al. (2005)	TOEFL iBT writing • Task: independent vs. integrative • Score levels: bands 3, 4, and 5 out of a 5-point scale	23 features categorized into lexical and syntactic complexity, grammatical accuracy, argument structure, orientations to evidence, and verbatim uses of source text	• Grammatical accuracy distinguishes among writing score bands 3, 4, and 5.
Biber & Gray (2013)	TOEFL iBT speaking and writing • Mode: speaking vs. writing • Task: independent vs. integrative • Score level: scores on a 4-point scale	171 lexico-grammatical features	• Four features were significantly more frequent in low-scoring independent speaking tasks • Many features had no significant relationship to the score • Most grammatical features were weak predictors of score level
Knoch et al. (2014)	TOEFL iBT writing • Task: independent vs. integrative • Score level: scores on a 5-point scale	24 features categorized into accuracy, fluency, syntactic and lexical complexity, coherence, cohesion, content, orientation to evidence, and metadiscourse	• The number of ideas in the essay, density of patterns (i.e., native-like formulaic structures), accuracy (i.e., error-free T-units and clauses), number of words strongly distinguished the score levels
Banerjee et al. (2007)	IELTS Academic writing • L1: Chinese vs. Spanish • Task: 2 different tasks • Score level: IELTS bands 3-8	Cohesive devices, vocabulary richness, syntactic complexity, grammatical accuracy	• Lexical richness and grammatical accuracy (on subject-verb agreement and passives) and discriminated well between the score levels

Table 1.2 (cont'd)

Banerjee et al. (2015)	ECPE writing • Score level: scores on a 5-point scale	Length, lexical diversity, lexical frequency, cohesion, syntactic complexity, prompt dependence	• Essay length and lexical knowledge were strong predictors • The variables predicted the score levels more accurately at the A, C, E levels, but less so the level in between
Yan & Staples (2016)	ECPE writing • Prompts: 3 different prompts • Score level: scores on a 5-point scale	41 lexico-grammatical features	• Five functional dimensions of lexico-grammatical complexity were found • Four dimensions (type of discourse, prompt dependence and lexical diversity, overt suggestions, and stance and referential discourse) correlated significantly with the holistic scores • Three dimensions were useful in predicting the levels

Note. TOEFL, Test of English as a Foreign Language; IELTS, International English Language Testing System; ECPE, Examination for the Certificate of Proficiency in English

In another large-scale English exam context, Banerjee, Francheschina, and Smith (2007) investigated the characteristics of writing performance on the IELTS Academic Writing section. Specifically, they investigated writing performance at IELTS bands 3 to 8 with respect to the use of cohesive devices, vocabulary richness, syntactic complexity, and grammatical accuracy. The researchers investigated 550 written texts produced by 275 examinees from two L1 groups, Chinese and Spanish. They found that all except the syntactic complexity measures investigated here were informative of increasing proficiency level. The researchers reported that grammatical accuracy such as subject-verb agreement and passives were good discriminators of the score level, that lexical density, variation and sophistication increased as the score increased, and the use of cohesive devices was more helpful in distinguishing between lower level bands.

Two studies on the *Examination for the Certificate of Proficiency in English* (ECPE) provide more focused research into the linguistic features and scores, with more sophisticated methods. Banerjee, Yan, Chapman, and Elliott (2015) analyzed previous studies that had utilized Coh-Metrix indices (Graesser, McNamara, Louwerse, & Cai, 2004) and selected the measures that significantly predicted writing proficiency and/or represented the following categories: text length, lexical diversity, lexical frequency, cohesion, syntactic complexity, and prompt dependence. They found that while text length and lexical knowledge measures were strong predictors of the score levels, syntactic complexity measures in Coh-Metrix were not useful in distinguishing between the levels. Based on a discriminant function analysis and classification using the select measures, the research team reported that the functions (i.e., the select Coh-Metrix measures grouped by discriminant functions) relatively accurately predicted the essays at the A, C, and E levels. However, they noted that these functions did not differentiate between the C and D levels. After triangulating the results with qualitative data analysis of rater discussions,

the researchers revised the rating rubric to include language that would be help raters better recognize the different characteristics shown in the examinee essays and better distinguish between the score levels.

Yan and Staples (2016) studied the co-occurrence patterns of 41 lexico-grammatical features, including grammatical and syntactic structures, semantic categories of grammar items (e.g., attributive adjectives, size adjectives, topical adjectives), different types of lexical bundles (e.g., prompt-matching, generic, stance), lexical features (e.g., type-token ratio, vocabulary frequency profile). Their aim was to find support for the ECPE's writing scale. Through a multidimensional analysis, the researchers found five dimensions underlying the investigated features. They also analyzed how the different dimensions correlated with the ECPE score and found that three dimensions significantly differed across score levels. The results suggested that essays receiving higher scores included more features associated with written language rather than spoken language, showed more lexical variety, and used more prepositions and referential discourse (e.g., *a lot of, more and more, the opportunity to*).

In summary, studies on the TOEFL and IELTS revealed that some measures of linguistic features differed depending on the scores given by the raters. Accuracy, lexical, and length measures were consistently found to be strong predictors of score level. However, these studies were concerned with overall validity of the test design (e.g., the effect of different types of tasks) and provided little discussion on the how the rating scale might be improved in relation to the learner language characteristics observed at different score levels (except for Knoch et al., 2014). Furthermore, while these studies presented significant findings on the differences in linguistic features across proficiency groups, they remained at a descriptive level and did not further investigate how the linguistic features might distinguish or predict the scores. The more recent

two studies on the ECPE, on the other hand, offered both description of the language performance by learners at different score levels and classification or prediction of the levels based on the use of the linguistic features, providing more support for the validity of the rating scale.

1.2 Studies on Grammar Use and Text Quality

There exists a large body of literature that explores the linguistic features of learner English in association with text quality (e.g., score on target writing or speaking performance). In this section, I first synthesize the research that has examined specific grammatical features (e.g., *that*-complement clauses) rather than the more traditional, automated indices of syntactic complexity (e.g., mean length of T-unit, number of dependent clauses) to find differences among varying score levels. While the studies described in the previous section were validation research to provide support for specific aspects of test validity (e.g., task type, parallel test form, rating rubric), the studies discussed in this section generally identify and describe the linguistic features characteristic of text quality. Though some of the studies were not conducted in a testing context, research that investigates the link between grammatical features and text quality using corpus-based methods provides insights into what characteristics of learner language a specific score may imply.

Table 1.3

Studies on Specific Linguistic Features in Relation to Score Levels

Study	Participants (L1)	Mode and tasks	Score levels	Investigated features	Exploratory analysis	Inferential statistics
Biber et al. (2016)	480 (various)	Speaking (TOEFL iBT) • Two independent tasks, four integrated tasks Writing (TOEFL iBT) • One independent task, one integrated task	1-4 1-4	23 features (word length, 22 grammatical items or structures)	Multidimensional analysis	Multifactorial ANOVAs
Friginal & Weigle (2014)	24 (various) and 51 (English)	Writing • TOEFL-style independent writing	High	23 features (3 length-related measures, 4 meta-discoursal elements, 4 parts of speech, 12 grammatical items or structures)	Cluster analysis	
LaFlair & Staples (2017)	98 (various)	Speaking (MELAB OPI) • Oral interview	7 score bands	41 features	Multidimensional analysis	
Jarvis et al. (2003)	120 (Arabic, Chinese, Spanish)	Writing • ESL placement test composition • Tests of Written English	High	22 features (2 length-related measures, lexical diversity, 4 meta-discoursal elements, 15 grammatical items or structures)	Cluster analysis	
Park (2017)	390 (Korean)	Writing • One narrative essay, one argumentative essay	High, mid, low	Two diversity measures based on the use of 13 verb-argument structures	Discriminant function analysis	Multiple linear regression
Taguchi et al. (2013)	54 (various)	Writing • Composition	High, low	15 grammatical features	Descriptive	

Table 1.3 (cont'd)

Yan & Staples (2016)[a]	595 (various)	Writing (ECPE) • ECPE writing section	1-5	41 features (31 grammatical and syntactic features, 4 measures of lexical bundles, 6 vocabulary features)	Multidimensional analysis	Correlations, factorial MANOVAs

Note. MELAB OPI, Michigan English Language Assessment Battery Oral Proficiency Interview

[a] This study was included in this table again to highlight the use of dimensional analysis followed by level prediction.

What most of these studies have in common, methodologically, is that they looked at several different lexico-grammatical features and performed a type of dimension reduction analysis to uncover patterns of feature co-occurrence (identified in the exploratory analysis column in Table 1.3). For instance, Jarvis, Grant, Bikowski, and Ferris (2003) and Friginal and Weigle (2014) identified the types of highly-rated L2 English learner essays depending on their use of a set of linguistic features by conducting a cluster analysis. Similarly, Biber, Gray, and Staples (2016), LaFlair and Staples (2017), and Yan and Staples (2016) measured the frequencies of various linguistic features to identify linguistic dimensions through a multidimensional analysis technique (Biber, 1988). These three studies further investigated how the resulting dimensions, or co-occurrences of features, predicted writing quality. Park (2017) offered novel measures of syntactic complexity based on English verb-argument constructions. She identified 39 verb-argument types in a corpus of written English and used the number of verb-argument types and the corrected type-token ratio of verb-argument constructions as syntactic variety and complexity measures. She found these measures to be strong indicators of proficiency, as they showed a high positive correlation with the proficiency levels and clearly discriminated between the different levels of proficiency. What these studies highlight is that co-occurrence patterns of linguistic features are more effective in characterizing proficiency levels than looking at how the use of an individual feature varies across levels (e.g., Biber & Gray, 2013).

1.3 Features of L2 writing: CALF Measures Across Score Levels

Syntactic complexity, accuracy, lexical complexity, and fluency (CALF) measures have been widely adopted as indices of L2 production used to examine language development, the effect of pedagogical methods (e.g., instruction, feedback; see Wolfe-Quintero et al. 1998) or

task complexity (see Johnson, 2017 for a research synthesis). In line with the focus of this study, I only summarize studies that have investigated the relationship between CALF measures (syntactic complexity included) for L2 written production and writing quality. As illustrated in Table 1.4, most studies have found that some measures of linguistic features predict or correlate with writing quality. Because many different combinations of measures have been used in this line of research, it is not easy to draw hard conclusions. However, length-related measures (e.g., mean length of clauses, mean length of sentences, mean length of T-units), types of lexical measures (e.g., lexical complexity, diversity, sophistication), and accuracy measures (e.g., error frequency) were consistently found to be predictive of or correlate with writing quality. New measures based on a usage-based approach utilizing a native English reference corpus, such as word association strength and word frequency, were also found to be predictive of writing quality (e.g., Bestgen & Granger, 2014; Kyle & Crossley 2017).

Interestingly, Thewissen (2013) found that a development in accuracy is most marked between the B1 and B2 levels, and that error rate showed stabilization through the higher levels (i.e., C1 and C2). The types of errors that showed improvement included lexical phrase errors, spelling errors, countable/uncountable distinction errors, preposition (noun-dependent and verb-dependent) errors, adjective errors, missing words, unclear sentences, and complex connector, morphology, pronoun and determiner errors. The error types that showed no developmental change across levels, which accounted for 35% of all error types, included tense errors, punctuation related errors, verb complementation, finite/non-finite forms, genitive, noun complementation, and adjective dependent prepositions. However, the author noted that except for punctuation, these linguistic features appeared infrequently.

Table 1.4

Studies on Various CALF Measures in Relation to Score Levels

Study	Participants/ Texts	Mode and task	Score levels	Investigated features	Findings
Alexopoulou, Michel, Murakami, Meurers (2017)	174,771 learners	Texts written on 6 tasks (narrative, descriptive, professional)	Englishtown level 1-16 (equivalent to CEFR A1-C2)	4 accuracy measures, 3 syntactic complexity, lexical complexity	• Relative error frequency decreased as proficiency advanced
Bestgen & Granger (2014)	57 learners, 171 essays	Descriptive essay	Writing scores on vocabulary and language use categories	CollGram types, CollGram association strengths, novel accuracy measures (absent tokens and types in native reference corpora)	• CollGram types and association strength, accuracy measures significantly correlated with language use and vocabulary scores
Bulté & Housen (2014)	45 learners, 90 texts	Descriptive essay	Holistic writing score, language use and vocabulary subscores	10 syntactic complexity measures, 3 lexical complexity measures	• 7 measures of syntactic complexity correlated with overall writing quality • 8 measures of syntactic complexity correlated with language use • one lexical complexity measure correlated with both writing quality and vocabulary subscore
Crossley & McNamara (2014)	57 learners	3 descriptive essays written over one semester	holistic, combined score	11 clausal and phrasal features	• Incidence of all clauses, infinitives, that verb complements significantly predicted rating

Table 1.4 (cont'd)

Kyle & Crossley (2017)	240 learners; TOEFL 480 essays	Writing; argumentative	5-point scale	Syntactic sophistication and complexity	• More strongly associated verb-VAC combinations and less frequent verb-VAC combinations, and MLC significantly predicted score level
Lahuerta Martínez (2018)	188 learners (L1 Spanish)	Argumentative essay	Lower intermediate (3rd grade, A2) and intermediate (4th grade, B1)	8 measures of syntactic complexity in 4 categories: MLS, sentence composition, proposition combining and clause linking, syntactic phrasal complexity	• 7 measures (MLS, compound and complex sentence ratios, coordinate and dependent clause ratios, nominalization) significantly correlated with holistic writing score • 5 measures (MLS, compound and complex sentence ratios, coordinate and dependent clause ratios) were significantly different between two levels
Lu (2011)	3,554 texts	3 genres (narrative, argumentative, expository)	Institutional level (1-4)	14 measures of syntactic complexity	• 7 measures showed significant changes between levels in a linearly increasing trend (3 length measures, 2 complex nominal measures, 2 coordinate phrase measures)
Matthews & Wijeyewardene (2018)	104 texts	IELTS writing task	10 score levels	57 measures in 6 categories: referential cohesion, connectives, lexical diversity, word information, syntactic complexity, syntactic pattern density	• Argument overlap, type-token ratio, and second-person pronouns significantly differentiated low and high score groups • Argument overlap index and second-person pronouns were significant predictors of the score level

Thewissen (2013)	223 learner texts (L1 Spanish, German, French)	argumentative essay	CEFR B1, B2, C1, C2	40 types of errors	• 17 types of errors showed differences between B1 and B2 levels (e.g., lexical phrase, countable/uncountable, unclear sentences, pronoun/determiner) • 16 error types did not show developmental patterns (e.g., punctuation, verb complementation, finite/non-finite forms)
Verspoor, Schmid, & Xu (2012)	437 texts (L1 Dutch)	Narrative writing	5 levels calibrated to CEFR (A1 – B1)	10 sentence-level complexity measures, 7 verb phrase measures, 7 types of chunks, 2 lexical sophistication measures, 8 types of errors	• Complex and compound-complex sentences increased at all levels • Sentences with dependent clauses increased at all levels • The Guiraud index (corrected type/token ratio) and frequencies of chunks increased at all levels • Total errors decreased between levels 1-3, 2-4, 3-5 and between 3-4

Notes. VAC, Verb-argument construction; MLC, mean length clause; MLS, mean length sentence.

For the purposes of interpreting test scores and validating testing instruments, investigation into linguistic features of written samples needs to make direct links to test design, rating materials, and test validity. First, as Xi (2017) and Egbert (2017) discussed, the linguistic features chosen for investigation should reflect the constructs being measured by the test and be clearly operationalized. For example, the use of complex nominal clauses, though considered a hallmark of academic writing, may not be an ecologically valid measure for a test that does not require academic writing and does not include such construct in the rating rubric. Secondly, investigation into the relationship between linguistic features, such as the interaction between syntactic complexity and accuracy, requires more research because human raters holistically consider various aspects of language use (Bulté & Housen, 2014). Especially regarding the CEFR scale, the CEFR regards range and accuracy as the two pillars of language use (Council of Europe, 2001, 2018). That is, both the degree of complexity and sophistication, and degree of accuracy are important constructs in this scale. As will be presented in Table 2.1, the CELC rubric descriptors also describe the degree of complexity and accuracy in the grammar and vocabulary categories. This is also true for not only the tests aligned with the CEFR (e.g., CEPE, Cambridge English Assessment), but also widely-used tests such as the IELTS and TOEFL. Therefore, the relationship between complexity and accuracy requires more investigation, especially with regard to the supposed linearity present in level descriptors, i.e., that there is a linear increase in complexity and accuracy as proficiency advances. Lastly, to inform language testing, research needs to examine how accurately the combinations of appropriately set target linguistic features classify learner-produced texts into score levels. As exemplified in Banerjee et al.'s (2015) study, the rating rubric descriptors need to be able to guide raters to differentiate between adjacent levels.

1.4 The Common Reference Levels

The Council of Europe (2001, 2018) described a framework of six standard language proficiency levels that range from basic to advanced: A1, A2, B1, B2, C1, and C2 in the *Common European Framework of Reference for Languages*. The document lays out the schemes, parameters, categories of use, and examples of the Framework. It also presents a variety of illustrative descriptors for different purposes and activities in the learning, teaching and assessment of languages: language domains, situations, types of communication, text genres, strategies and learners' language competences. For instance, the CEFR document provides the illustrative descriptors for the general linguistic range at each proficiency level (for the full description, see Council of Europe, 2001, p. 110).

Since its release, the CEFR has received much criticism in many regards (see Deygers, Zeidler, & Vilcu., 2017). With respect to the scale descriptors, Hawkins and Buttery (2010) critically pointed out that the descriptors fail to mirror the varying range of grammatical constructions and lexical items that learners at a specific level can or cannot produce. Since then, researchers have undertaken the mission to improve the existing descriptors using empirical data under the project titled *English Profile*. The English Profile is a series of studies investigating authentic learner English at different proficiency levels by employing the Cambridge Learner Corpus, which consists of written English texts from the Cambridge ESOL examinations.

1.4.1 Grammatical features of CEFR levels

One of the projects undertaken by the English Profile research program specifically pertains to identifying grammatical features that are characteristic of each of the CEFR proficiency levels, driven by the notion that "there are certain linguistic properties that are characteristic and indicative of L2 proficiency at each level (Hawkins & Filipović, 2012, pp. 5-

6).” The term *criterial features* reflects this very idea as well as the assumption that there exists a set of linguistic features that help differentiate one level from another. Hawkins & Filipović (2012) reported their search for the grammatical criterial features in a publication entitled *Criterial Features in L2 English*. It includes a list of linguistic structural features for each CEFR level, from A2 through C2, that were attested in a subset of the Cambridge Learner Corpus, a corpus which includes texts from the Cambridge Main Suite of General English (i.e., five exams for CEFR levels A2 through C2). The grammatical criterial features were strongly informed by (a) the series of studies by Ek and Trim (1990a; 1990b; 2001, as cited in Hawkins & Filipović, 2012) on the description of language proficiency, and (b) the patterns of verb co-occurrence at different CEFR levels identified by Williams’s (2007, as cited in Hawkins & Filipović, 2012) study. Based on their observation of the Cambridge Learner Corpus, Hawkins and Filipović themselves additionally specified four types of raising constructions, which were considered to be complex syntactic structures acquired at later stages of language learning, as well as four types of double embedded genitives.

Except for the list of these sources, there is little information on the actual procedures followed in searching for, identifying, and selecting the criterial features. It appears that once Hawkins and Filipović had pre-selected a list of structures, they obtained the target feature frequencies in the corpus. They explained that the criteriality, i.e., to which level each feature belonged, was determined by what is called the 10-to-1 rule: If the ratio of the number of occurrences for feature X was higher than 10 to 1 in favor of a given level, X was considered a criterial feature of that level (Hawkins and Filipović, 2012, p. 37). For example, the authors observed that subject-to-subject raising constructions with the verb *seem* (e.g., “John *seems* to be a nice guy.”) occurred 3 times at the A2 level and 46 times at the B1 level in the corpus.

Therefore, they categorized the construction as criterial to B1. While the authors acknowledged that this rule is not without problems, they explained that it worked as a practical and consistent way of determining the criteriality of the features.

O'Keeffe and Mark's (2017) ongoing research also employs the Cambridge Learner Corpus to investigate what learners at different CEFR levels can do with grammar. Their approach differs from the grammatical criterial features study in that they refer to a list of grammar items that are perceived as having to be instructed at different levels of proficiency. These researchers thus gathered grammar syllabi of ELT pedagogical grammars and course books and categorized the grammatical items that appeared in these sources. Their pilot study investigated how these items were attested in the Cambridge Learner Corpus. Important findings from this pilot study include: (a) there was low probability that a grammatical form and a level would correspond one-to-one, (b) employing more lexis goes in tandem with an expanding repertoire of grammatical structures, and (c) the developmental pattern echoes stabilization, where a syntactic form is mastered at a certain level then takes on a greater complexity of meaning and is employed with greater command at the higher levels. One illustrative example they provided was the use of the "adjective + *that*-clause" construction by learners at different proficiency levels: "I *am sure (that)* we will find something to do (A2)"; It *seems obvious that* this oil comes from the gas station (B2)"; "It *is highly unlikely that* the goods can vanish from your warehouse (C1)".

In short, corpus-informed research has shown that grammatical features can provide valuable and concrete descriptions of what language learners can do linguistically at different CEFR levels. However, examining more data from different corpora would help further substantiate and solidify how such features are used by English learners.

1.4.2 Previous studies focusing on the CEFR grammatical features

The English Profile studies have presented empirical findings on various linguistic properties at each CEFR proficiency level, drawing on a large corpus representative of diverse L1 backgrounds. However, attempts to empirically examine and extend their findings in contexts beyond the Cambridge Learner Corpus have been scarce. To my knowledge, Juknevičienė and Šeškauskienė (2014) were the first to examine the list of grammatical criterial features identified by Hawkins and Filipović (2012) outside the Cambridge Learner Corpus. They examined the validity and applicability of these features by drawing on a corpus of Lithuanian learners of English. The corpus consisted of 433 essays written for a college entrance exam that received passing scores (above 60 out of 100). The researchers analyzed the raw frequency of 49 target structures that occurred in the corpus. They reported that 27 of the features either did not occur at all or occurred very infrequently. About half of the A2-level features and about one third of B1 and B2 structures were more commonly used (i.e., occurring more than 50 times). The researchers interpreted the results from the perspective of L1 transfer, explaining, for example, that the high frequency of the structure adverbial subordinate clauses with *-ing* (e.g., "I waited for the train, reading a novel.") could be attributed to the frequent use of participial forms in Lithuanian verbs. The researchers thus argued that one should not expect the criterial features to be readily applicable to all contexts. Though they reported only raw frequencies and no other measures or statistics, the findings from this study hint at the need to expand the research context to incorporate diverse settings and learners with different L1s.

1.5 Purpose of the Study

The goal of this study is to expand the previous research on the relationship between learner language and proficiency levels. In this study, I conduct an empirical, corpus-based

analysis of grammatical features in a new testing context (i.e., the *Certificate of English Language Competency*) by examining features identified to be relevant to CEFR levels (Hawkins & Filipović, 2012; English Profile, 2015) and constructs of linguistic competency (i.e., linguistic range and control) described in the CEFR (Council of Europe, 2001; 2018). At the local level, this study will make direct and practical contributions to improving the CELC scale descriptors crucial for rating processes (e.g., rater training and use of the rating rubric). The empirical examination into the relationship between the test's rating scale, scale descriptors, and the features of the texts produced by the examinees, however, will have broader implications for language test validation research because all performance tests include these three components. This study also contributes to the effort to provide more concrete descriptions and profiling features of learner English at specific proficiency levels, by examining a rare corpus of learner that is tied to a well-established proficiency scale (i.e., CEFR). In summary, I offer (a) a demonstration of finding evidence to support the validity and usefulness of the rating scale and descriptors in the rubric, thus contributing to test validation research; (b) an empirical investigation into the use of the grammatical features of the CEFR levels, which has implications for understanding learner English as related to the CEFR levels, which can further serve as a great resource for any educational context that utilizes the CEFR. With these purposes in mind, I have the following research questions to guide the present study:

1. To what extent do the CELC examinees use the grammatical features identified as indicative of CEFR levels? What are the co-occurring patterns of these grammatical features?

2. How well do the co-occurring patterns predict the examinees' writing performance as measured by rater-assigned grammar score?

3. Is there any interaction between the patterns of grammatical features and accuracy in

predicting grammar scores?

4. How do the findings support test validity with regards to descriptions of grammatical ability in the rating rubric across the score levels?

a) Do the CELC texts produced by examinees reflect the descriptions of grammatical performance in the rating rubric?

b) Do the linguistic features of examinees' texts differentiate between score levels?

In addressing the last research question, I specifically examine the following assumptions adopted from Knoch and Chapelle (2018), which address the validity of the rating rubric descriptors and scale.

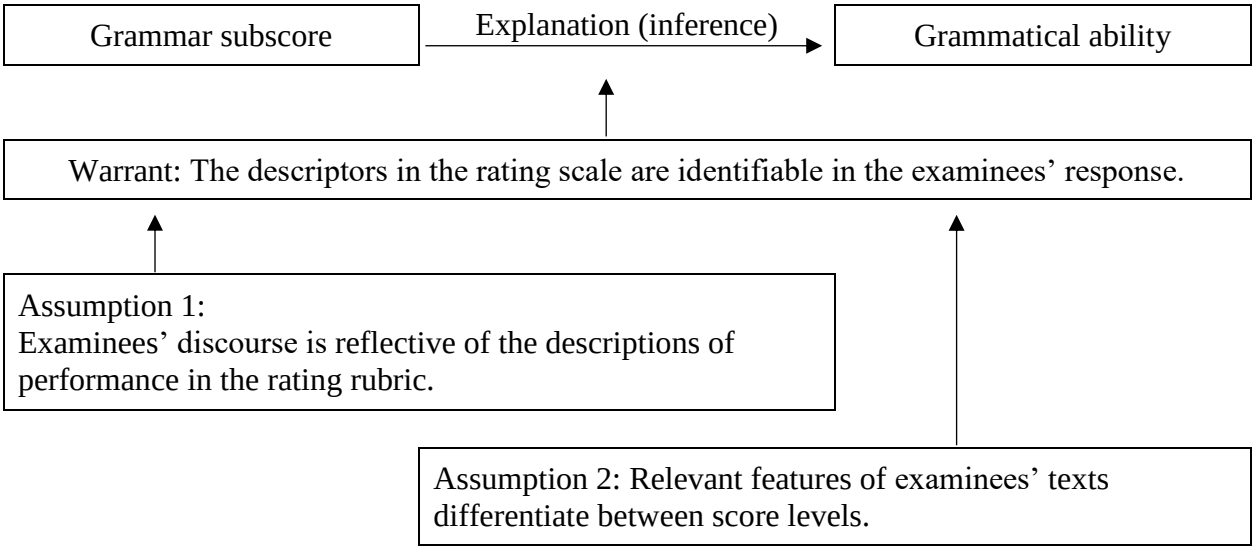

Figure 1.2. *Argument-based framework for the current study.*

For the grammar subscore to validly explain the examinees' grammatical ability, it needs to be warranted that the descriptors in the rating scale are identifiable in the examinees' response, thereby correctly explaining the meaning of the score. The relationship between the descriptors in the rating scale and the examinees' response can be explicated by examining (a) whether what is described in the rating rubric actually exists in the examinee responses (Assumption 1) and (b)

whether characteristics of examinee responses reliably distinguish between score levels (Assumption 2). That is, if research comparing linguistic features across score levels finds distinctive features for each level, this would serve as evidence for scale validity. Finding such evidence in the actual examinee texts is important both in terms of supporting the overall validity argument of the test and substantiating the effectiveness of the rating scale and descriptors that are utilized in test operations (Knoch et al., 2014).

CHAPTER 2: METHODOLOGY

2.1 Context of the Study: Certificate of English Language Competency (CELC)

The *Certificate of English Language Competency* (CELC) is a four-skill exam developed by the English Language Center Testing Office at Michigan State University (Michigan State University English Language Examinations, n.d.). It is a large-scale assessment administered throughout Greece with approximately 3,000 exam candidates per each year. The exam is designed to assess English language ability at CEFR Level B2 in all four modalities (i.e., listening ability, speaking ability, reading ability, and writing ability). Each section is worth 25 points, and candidates who earn an overall score of at least 60 points out of 100 possible points obtain a certificate and a score report. There are several large-scale exams that certify B2-level English proficiency, including: two versions of *B2 First* and *B2 Business Advantage* (BEC Advantage) by Cambridge Assessment, the *Pearson Test of English General* (Level 3) by Pearson, and the *Examination for the Certificate of Competency in English* by Michigan Language Assessment. Demonstrating B2-level competency is significant in that this is the level most commonly used level in meeting the language requirement for university entrance (Deygers et al., 2017). As the B2 level represents a high-intermediate level, the examinees and score levels

for B2 level certification are potentially inclusive of a range of proficiency, from beginner to advanced. However, it should be noted that although the test's rating scale is associated with score levels that range below and above the B2 level, neither CELC scores nor CELC score levels equate to other CEFR levels.

2.1.1 Writing task and scoring

The data used in this study consists of examinee essays written for the writing section of the exam, which includes a writing task (20 points) and a multiple-choice grammar test (5 points). The writing task requires examinees to write an opinion essay in response to one of the two prompts offered. Here is an example set of prompts:

- Imagine you found a bag of money under the bushes on a street near your house. What would you do with the money? Why?
- Teachers at your school are asking students about ways to make the school better. In what ways could your school be better? What changes need to be made?

Examinees are given 35 minutes to complete the task. There is no specified word limit. They are instructed to write as much as they can, and to provide supporting examples, reasons, and explanations. All essays are hand-written at the time of the exam. Two raters rate each essay, using an analytic rubric with four categories: grammar, vocabulary, task completion, and genre appropriateness (see Table 2.1 for grammar subcategory rubric descriptors, and Appendix A for the complete rubric). The subscore for each category ranges from 1 to 5, and the possible maximum score for the essay is 20. If there is a discrepancy greater than 3 points between the two raters, a third rater adjudicates, and the essay score is determined by averaging the two closest scores among the three.

2.1.2 Rating materials

The CEFR's interpretation of language competences consists of four types: strategic competence, linguistic competence, pragmatic competence (comprising both discourse and functional/actional competence), and socio-cultural competence (Council of Europe, 2018). The descriptors for linguistic competence can be categorized into three types: *Range*, *Control*, and *Phonological and Orthographic Control*. The rating rubric for the CELC reflects these key aspects of linguistic competence, with descriptors that illustrate syntactic range and control (i.e., accuracy).

Table 2.1

Grammar Category of the Rating Rubric for CELC Writing (highlights made by the author)

CELC Rating Scale	**Linguistic Competency - Grammatical Accuracy**
5: Honors Pass (exhibits some C1/C2 features)	Meets all B2 requirements for this category, plus effectively uses advanced structures such as **multi-clausal sentences** and **syntactic variety** to effectively clarify, explain and elaborate support for the point of view assumed in the essay
4: Clear Pass (exhibits B2/C1 features)	Control of **a range of syntactic forms** that allows writer to efficiently and effectively convey meaning and ideas relevant to the B2 task; **few errors**
3: Marginal Pass (B2 "floor")	Control of **basic syntactic forms is adequate** to convey meaning and ideas relevant to the B2 task without causing confusion, even though **some errors** may be present
2: Narrow Fail (satisfies some, but not all B2 criteria)	Control of **basic syntactic forms is NOT adequate** to convey meaning and ideas relevant to the B2 task without causing confusion; **numerous errors** are present and **limit effectiveness of the text**
1: Fail	Telegraphic, severely limited, may be rudimentary or unintelligible

The descriptors for the grammar category highlight how effectively an examinee can convey his or her intended message by manipulating syntactic forms of the target language. The descriptors relevant to grammatical range depict a progression from "control of basic syntactic forms" that are "not adequate" and "limited" to effective use of "advanced structures such as multi-clausal sentences and syntactic variety." Therefore, the descriptors allude both to examinee

ability to use various syntactic forms, and to the degree of accuracy of those forms. Since the descriptors can be interpreted subjectively and relatively, raters are advised to refer to essay benchmarks in rater training. The CELC developers provide sample essay benchmarks with annotations on their webpage for examinees as well. Examples at each score level are provided in Table 2.2. The evaluative comments (in the third column) refer to the excerpt from the benchmark essays (in the second column).

Table 2.2

CELC Benchmark Essays and Comments on Grammar

Score level	Annotated text from benchmark essays	Evaluative Comments
5	"For instance, I could tell him that what he's doing is no good for himself and one good argument for this, was that he would not be able to copy in the final exams and he would have to repeat the class."	• This essay uses advanced structures such as multi-clausal sentences and syntactic variety, in spite of some spelling errors and minor grammatical errors.
4	"When a person need help for his homework or for his job I think that someone that he knows him will help him."	• The writer displays a range of syntactic forms (e.g., simple sentences, compound sentences, complex sentences). • Some grammatical errors in subject-verb agreement (e.g., "a person need") and resumptive pronouns (e.g., "someone that he knows him").
3	"Writing about the theme of my friend's copying homework from the time that I had said to him that I will help him to finish him homework, but of course not to copy mine."	• The writer displays control of basic syntactic forms, but displays some distracting errors
2	"I believe is not true for my friend to copy my homework but is my friend and must be good with him" "First of all I ask my friend why do this and he told the truth because is my friend and if he do this, he do not there are some advice for this."	• Persistent grammar errors (e.g., subject dropping, lack of subject-verb agreement) are distracting, limit the effectiveness of the text… • Basic syntactic forms are evolving but are still not adequately controlled to convey meaning and ideas.
1	"One days my call for a friends and I asked me for a help."	• Very rudimentary control of even basic syntactic forms renders the essay almost unintelligible.

The evaluative comments refer to the rating rubric, to explain and justify which aspects of the essay led the raters to assign a specific score. To investigate whether the rating rubric descriptors truly characterize the texts produced by the examinees, I examined examinees' use of numerous types of clausal-level constructions and the linguistic accuracy of their texts. For this purpose, I built a corpus of essays, described in the next section, produced during CELC administrations.

2.2 Data for the study

2.2.1 Text selection

A total of 560 examinee texts from two CELC administrations (Spring 2016 and Spring 2017) were sampled based on the rater-assigned grammar subscore. The texts were written on one of the following topics: (a) comparing spending money to buy something that can be enjoyed right away versus saving money for a longer period of time to buy something more expensive, (b) the most important characteristic of a good friend, or (c) what teachers can do to make learning more fun. The numbers of essays written on each topic were comparable: 1,005 on topic A; 1,025 on topic B; and 1,304 on topic C. The overall distribution of the score levels is presented in Table 2.3, with the number of essays and percentage out of each topic in parenthesis.

Table 2.3

Essay Distribution per Grammar Subscore from CELC Spring 2016 and 2017

	Grammar subscore										
Topic	**0**	**1**	**1.5**	**2**	**2.5**	**3**	**3.5**	**4**	**4.5**	**5**	**Total**
A	0	4	12	210	209	316	150	80	16	8	1,005
	(0.0)	(0.4)	(1.2)	(20.9)	(20.8)	(31.4)	(14.9)	(8.0)	(1.6)	(0.8)	
B	1	10	23	172	209	307	152	116	22	13	1,025
	(0.1)	(1.0)	(2.2)	(16.8)	(20.4)	(30.0)	(14.8)	(11.3)	(2.1)	(1.3)	
C	1	21	39	290	248	394	156	130	21	4	1,304
	(0.1)	(1.6)	(3.0)	(22.2)	(19.0)	(30.2)	(12.0)	(10.0)	(1.6)	(0.3)	
Total	2	35	74	672	666	1017	458	326	59	25	3,334
	(0.1)	(1.0)	(2.2)	(20.2)	(20.0)	(30.5)	(13.7)	(9.8)	(1.8)	(0.7)	

One notable characteristic of the CELC essays was that they were heavily concentrated within grammar subscore of 2 and 3 (70.7% altogether). This trend is visible in Figure 2.1, which presents the score distribution of the texts and the relationship between mean grammar score and overall essay score.

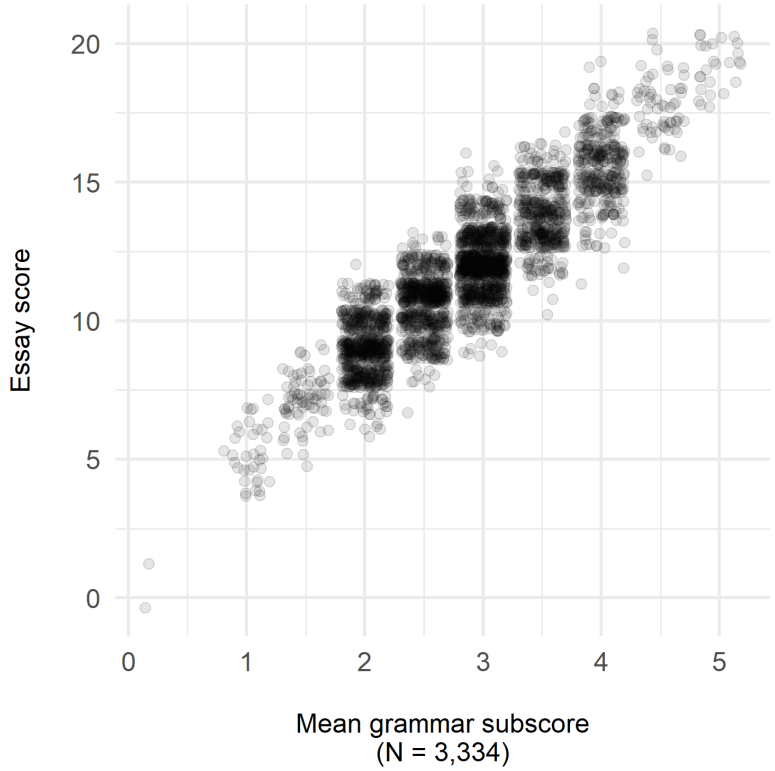

Figure 2.1. *Distribution of essays by mean grammar subscore and essay score (out of 20).*

As this study focuses on the differentiation between score levels, I selected only essays that received the same grammar subscore from both raters. That is, if an essay received a 2 from one rater and a 3 from another, it represents neither score level 2 nor 3. I also excluded essays that received an average grammar subscore of 1 because they were very rare, very short in length, and often unintelligible. I examined essays that received an average score of 1.5 (i.e., one rater assigned a score of 1 and the other a score of 2) and treated them as score level 1. Similarly, because the number of texts that scored 5 were very few, essays receiving an average score of 4.5 (to which one rater assigned a score of 4 and the other a score of 5), along with those receiving score 5 from both raters were considered to be at the highest score level. From here on, I refer to the grammar subscore in this selected corpus as score level. To reduce the influence of a specific topic, I sampled a roughly equal number of essays on each topic within each score level.

Table 2.4

Grammar Subscore Distribution of Selected Texts

	Grammar subscore					
Topic	**1-1.5**	**2**	**3**	**4**	**4.5-5**	**Total**
A	12	45	50	45	22	174
B	20	45	50	45	24	184
C	22	47	58	56	19	202
Total	54	137	158	146	65	560

I typed the selected essays into text files and corrected minor spelling mistakes (e.g., *expereience* to *experience*), corrected homonym errors only when the context was clear (e.g., *by* instead of *buy*), and changed British spellings to American (e.g., *organise* to *organize*). These corrections were made so as to not impede the accuracy of the natural language processing tools that automatically parse language. However, when coding for errors, the original texts without corrections were used so that the coders could capture all error instances.

2.2.2 Target grammatical features

Grammatical features were selected for examination based on Hawkins and Filipović's (2012) study on criterial features of the CEFR English levels and were refined by consulting a more detailed and updated version of the *English Grammar Profile* (English Profile, 2015). The motivation was to utilize linguistic features that were relevant to the test context and construct (i.e., CEFR level certification and grammatical ability). I selected complex clause-level features from the list provided by Hawkins and Filipović (2012), and narrowed them down to those that tend to occur frequently and have been included in a number of existing L2 writing studies in order to make comparisons to previous research (e.g., Biber, Gray, & Poonpon,2011; Biber & Gray, 2013; Biber et al., 2016; Staples & Reppen, 2016).

Features from the English Grammar Profile. Table 2.5 summarizes the features, each accompanied by a brief description and example along with its associated CEFR level. From here on, the features are referred to by the associated number presented in this table (e.g., the label F1.1 is used to refer to post-nominal modification with relative clause). The CEFR level indicates the level at which the use of a given grammatical feature becomes typical (as defined by the 10-to-1 rule explained in Section 1.4.1), according to the *English Grammar Profile*. Note that in some cases, one structure is considered as characteristic of two different levels depending on its lexis. For example, if a more common verb is used for item F3.1 (e.g., I *want* to go fishing), it is considered a B1-level feature, while if the verb is less common (e.g., she *failed* to finish the work), it is considered a B2-level feature.

Features from corpus-based studies on grammatical features. Biber et al. (2011) presented a critical view on using T-unit measures and dependent clauses (e.g., mean length of T-unit and clauses per T-unit) for assessing L2 writing development. They argued that these

measures were not suitable for capturing complex grammatical features and proposed to investigate specific grammatical devices which were categorized into three grammatical types: finite dependent clauses, non-finite dependent clauses, and dependent phrases (non-clausal). They analyzed the use of 28 specific grammatical features at the phrasal and clausal level to differentiate between registers (i.e., spoken vs. academic writing). All or parts of this set of grammatical features have been used in a number of studies since then (e.g., Biber & Gray, 2013; Biber et al., 2016; Staples & Reppen, 2016; Staples, Egbert, Biber, & Gray, 2016). Most of the features that are investigated in the current study were included in the list by Biber et al. (2011): relative clauses (F1.1), *that*-complement constructions (F2.1 – F2.3), *to*-complements constructions (F3.1 – F3.6), *wh*-word clauses (F4.1, F4.2).

Table 2.5

Target Grammatical Features

Feature description	**Example sentence**	**CEFR level**
1. Post-nominal modification with		
1.1 Relative clause	…borrowing things ***that*** you probably don't need. my friend ***who*** *is from Thessaloniki* introduced me…	A2/B1
2. *That*-complement clauses controlled by		
2.1 verb	we ***predict*** *(that)* water is here	B1
2.2 (*it* extraposed) adjective	it is ***strange*** *(that)* he went there	B1
2.3 (*it* extraposed) noun	it is my ***belief*** *(that)* we need to…	B1
3. *To*-complement clauses controlled by		
3.1 verb (subject-to-subject raising)	she ***failed to finish*** the task…	B1/ B2
3.2 verb (subject-to-object raising)	I ***found*** *the task* ***to be*** difficult	C1
3.3 verb (passive)	They are asked ***to be*** there …	B2/C1
3.4 adjective (it extraposed)	It is so ***easy to find*** news about people…	B1/B2
3.5 noun	We have ***money*** *to give* him.	B2

Table 2.5 (cont'd)

4. *Wh*-word clauses (as subjects or objects)		
4.1 Wh-word + to-complement	He told me ***what*** *to do.*	B2
4.2 Wh-word pseudocleft (WH-NP-VP)	***What*** *he had don*e was unclear.	B1
5. Ditransitive clauses		
5.1 Ditransitive (NP-V-NP-NP)	I ***bought*** *my grandma a present.*	A2
5.2 Prepositional dative (NP-V-NP-PP)	My friend ***gave*** *it to me.*	A2

Description of the target grammatical features. The following feature description are largely based on Hawkins and Filipović (2012, pp. 114-127).

- Feature category 1: Relative clauses.

  This category included *that*-relative clauses and *wh*-word relative clauses. I included both "relative pronoun + NP + VP" (e.g., the girl *that* I like) and "relative pronoun + VP" (e.g., my friend *who* likes hip hop) constructions.

- Feature category 2: *That*-complement clauses.

  The syntactic features in this category include finite *that*-clause complements. F2.1 consists of verbs with a finite complement clause: NP-V-S. I also included verbs with objective and prepositional phrases, such as: "He *told me that*…" and "He *said to me that*…." F2.2 and F2.3 refer to *it* extraposed structures with *that*-complement clauses. F2.2 is controlled by adjectives, as in "*It is strange (that)* he went there." F2.3 is controlled by nouns, as in "*It is a pity (that)* it happened," or "*It is my belief (that)* we need to save money for the future." All features with *that*-complement clauses can have either overt complementizer *that* or zero *that*.

- Feature category 3: *To*-complement clauses.

  These features include subject-to-subject raising verbs and adjectives and subject-to-object raising verbs and adjectives. Subject-to-subject raising English structures, for

instance, “She *failed to finish the task*,” are those in which the subject of the subordinate clause (i.e., *she*) has been raised out into the subject position of the main clause. “She” is the logical subject of “finishing the task,” but the whole event, not *she* herself, is claimed to have “failed.” In F3.2 constructions, subject-to-object raising verbs, “I *found the task to be difficult*,” the subject of the subordinate clause (i.e., the task) has been raised out and placed into the object position of the main clause. “The task” is the logical subject of “being difficult” and not semantically related to the verb “found.” These types of structures are also possible with passive verb, such as “They *were asked* to be there.” Here, the subject of the subordinate clause has been raised into the higher object position (i.e., We asked *them* to be there), then further promoted to the subject position in the main clause with the use of the passive construction.

- Feature category 4: *Wh*-word clauses.

  F4.1 is another *to*-complement clause with a *wh*-word as its head (e.g., He told me *what to do.*) and serves as a subject or object of a sentence. F4.2, *wh*-word pseudocleft, has the structure of WH + NP + VP (e.g., *What he had done* was unclear.) as subject or object. With this type, the *wh*-word is a direct object in the clause (i.e., *what* is the object of *he had done*).

2.2.3 Procedure

In this section, I provide details of the steps I took in corpus preparation, annotation, and search for target features in the corpus. I used the programming language *R* (R Core Team, 2019) and two packages for natural language processing and analysis to perform automatic annotation and feature search. The overall process is illustrated in Figure 2.2.

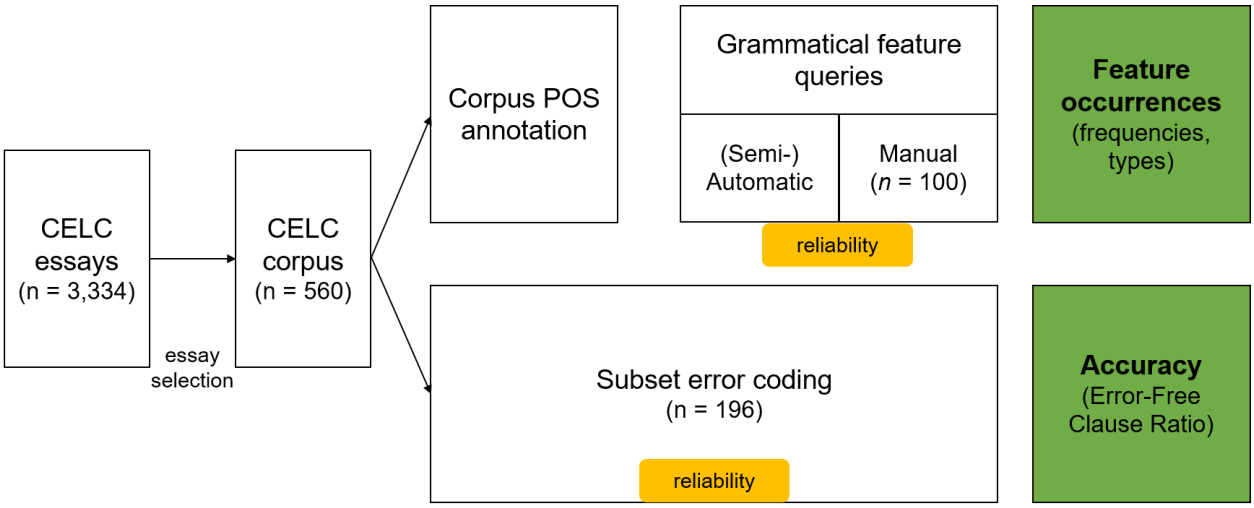

Figure 2.2. *Corpus preparation, annotation, and search procedure.*

Automatic annotation. The UDPipe (Straka & Straková, 2017) universal dependency analyzer was used via the *R* package *cleannlp* (Arnold, 2018) for language processing. The tool tokenizes raw texts, parses them into the base forms of words (i.e., lemmas), and automatically annotates parts of speech (POS; see Appendix B for a list of the tags). The resulting annotations are illustrated in Figure 2.3, where each row contains information about each word. Though the POS tagging is reported to have a 93.5% accuracy rate (Straka & Straková, 2017), its precision with learner language is largely unknown. Therefore, in this study, the reliability of the search results was also examined (see Table 2.8).

id	Text
29219	This way can lead people to choose something inexpensive that can be enjoyed right away.

↓

id	sentence id	token id	word	lemma	pos	tag
29219	4	1	This	this	DET	DT
29219	4	2	way	way	NOUN	NN
29219	4	3	can	can	AUX	MD
29219	4	4	lead	lead	VERB	VB
29219	4	5	people	people	NOUN	NNS
29219	4	6	to	to	PART	TO
29219	4	7	choose	choose	VERB	VB
29219	4	8	something	something	PRON	NN
29219	4	9	inexpensive	inexpensive	ADJ	JJ
29219	4	10	that	that	PRON	WDT
29219	4	11	can	can	AUX	MD
29219	4	12	be	be	AUX	VB
29219	4	13	enjoyed	enjoy	VERB	VBN
29219	4	14	right	right	ADV	RB
29219	4	15	away	away	ADV	RB
29219	4	16	.	.	PUNCT	.

Figure 2.3. *Raw text (top) and parsed text (bottom).*

Feature search. To extract the number of target features, I used the *R* package *corpuslingr* (Timm, 2018), which enables searches for grammatical constructions and complex lexical patterns. It takes a search syntax which consists of a combination of word forms, lemmas, and POS to generate the search results in concordance format (see Table 2.7) or as a summary of data. For instance, I used the search syntax `NP (ADV| NEG)? (ADJ)? (a list of relative pronouns) NP` (which translates into a sequence of syntax and tag for a noun phrase, optional adverb or negation, optional adjective, a list of relative pronouns, and a noun phrase) to find one type of finite relative clause (e.g., I have a *dog that my friend gave me*). For each target feature, I constructed search syntax(es) that would retrieve the desired results (listed in Table 2.6). I manually annotated the target features in 100 randomly selected texts and compared the annotations against the search output to ensure high recall accuracy. Table 2.6 lists the search syntax I used to generate concordances of target features.

Table 2.6

Search Syntax for Grammatical Features

Feature description	Search syntax
1.1 Relative clause	• NP[a] (ADV|NEG)? (ADJ)? (relative pronouns) NP • NP (ADV|NEG)? (ADJ)? (relative pronouns) (MD)? (ADV)? VERB • NP NP (MD)? (AUX)? (ADV)? VERB
2. *That*-complement clauses controlled by	
2.1 verb: verb + *that*-clause	• VERB (NEG| ADV| out) THAT~IN • (a list of verbs[b]) (NEG| ADV| out)? NP (ADV)? (MD| VERB) • VERB (out)? (to)? NP (NEG)? (THAT~IN)?
2.2 adjective: *it* extraposed adjective + *that*-clause	• IT~PRP (MD)? (NEG| ADV)? BE~VERB (NEG| ADV)? ADJ THAT~IN • IT~PRP (MD)? (NEG| ADV)? BE~VERB (NEG| ADV)? ADJ NP (MD| VERB)
2.3 noun: *it* extraposed noun + *that*-clause	• IT~PRP (MD)? (NEG| ADV)? BE~VERB (NEG| ADV)? NP THAT~IN • IT~PRP (MD)? (NEG| ADV)? BE~VERB (NEG| ADV)? NP NP
3. *To*-clauses controlled by	
3.1 verb (subject-to-subject raising): verb + *to*-infinitive	• VERB (out)? (NEG| ADV)? to (NEG| ADV)? VB
3.2 verb (subject-to-object raising): verb + NP + *to*-infinitive	• VERB NP (ADV)? (to)? (NEG| ADV)? VB
3.3 verb (passive): *be* verb + verb-*ed* + *to*-infinitive	• BE~VERB (NEG|ADV)? (VBD| VBN) (NEG| ADV)? to (NEG| ADV)? VB
3.4 adjective (subject-to-subject raising): adjective + *to*-infinitive	• IT~PRP (MD)? (NEG| ADV)? BE~VERB (NEG| ADV)? ADJ (NEG| ADV)? to (NEG| ADV)? VB
3.5 adjective (subject-to-object raising): adjective + NP + *to*-infinitive	• ADJ for NP to (NEG| ADV)? to (NEG| ADV)? VB
3.6 noun (subject-to-object raising): NP + *to*-infinitive	• NOUN (NEG| ADV)? to (NEG| ADV)? VB

4. Wh-word clauses (as subjects or objects)	
4.1 WH-to-complement: wh-word + to-infinitive	• `(list of wh-words`[c]`) (NEG|ADV)? to VB`
4.2 WH-NP-VP: *wh*-word + NP + VP	• `(list of wh-words`[d]`) (ADJ|NN)? (IN)? NP (NEG|ADV)? (MD)? (NEG|ADV)? VERB`
5. Ditransitive clauses	
5.1 Ditransitive: verb + NP + NP	• `(predefined list of verbs 1`[e]`) NP NP`
5.2 Prepositional dative: verb + NP + prep. + NP	• `(predefined list of verbs 2`[f]`) NP (to| for| in) NP`

[a] NP is defined as: optional combination of determiners, adjectives, quantifiers plus a noun; pronoun plus pronoun (e.g., *he or she*)
[b] verb list from the search results that included *that*
[c] list of wh-words for F4.1: *who, what, where, when, how*
[d] list of wh-words for F4.2: *who, what, which, where, when, why, how, whoever, whatever, whichever, wherever, whenever*
[e] list of verbs for F5.1: *give, show, tell, cause, buy, lend, bring, call, ask*
[f] list of verbs for F5.2: *give, show, tell, buy, lend, bring*

I refer to the feature search as semi-automatic because unwanted results had to be manually removed. Table 2.7 shows two examples of a subject-to-object raising construction (F3.2) and two examples of a relative clause (F1.1) found in the CELC corpus. While both results for the raising construction fit the search syntax `VERB NP (to)? (NEG|ADV)? VB`, the second example, "*get more time to save*" was removed because it is not a raising construction. The first example of a relative clause, "*if he has them which I want*" is clearly ungrammatical. However, it does contain, or at least attempt, the target structure (i.e., "*if he has the characteristics that I want*"). Therefore, both examples were considered relative clauses.

Table 2.7

Example Search Results for Verb Raising Constructions

Target feature	**Number**	**Lemma**	**KWIC**
verb subject-to-object raising	Example 1	lead people to choose	way can **lead people to choose** something inexpensive that can be enjoyed right away .
	Example 2	get more time to save	In this way , people have **got more time to save** enough money for spending what they want without limits .
relative clause	Example 1	they which	will do more easily someone my friend if he has **them which** I want .
	Example 2	the people who	We have **the people who** want save their money for a short time

To identify *that*-complement clauses controlled by verb with zero *that*, I created a list of words that took a *that*-clause in this corpus and also appeared in the concordance. For example, the first search syntax for F2.1 (in Table 2.6) yielded a list of verbs that take *that*-complement clauses in this corpus. I performed the next search to find the structure with zero *that* using this list of features that I generated because without the word *that* performing as a signpost, it is challenging to identify *that*-complement clauses. Similarly, the automatic search results found a large proportion of unwanted results for ditransitive clauses. Therefore, based on the results from

100 manually annotated texts, I searched for a select list of verbs followed by noun phrases, which included: *give*, *show*, *tell*, *cause*, *buy*, *lend*, *bring*, *call*, and *ask*.

***Reliability of the feature search*.** To account for the reliability of the query results retrieved by the tagging and querying processes, I examined the agreement between (manually filtered) semi-automatic search results and manual annotation.

Table 2.8

Agreement Between Feature Coding Methods (n = 100)

Feature description	**Manual coding (count)**	**Semi-automatic search (count)**	**Recall rate**
1. Post-nominal modification with			
1.1 Relative clause	675	641	.95
2. *That*-complement clauses controlled by			
2.1 verb	507	443	.87
2.2 adjective	31	28	.90
2.3 noun	7	7	1.00
3. *To*-complement clauses controlled by			
3.1 verb (subject-to-subject raising)	304	283	.93
3.2 verb (subject-to-object raising)	115	108	.94
3.3 verb (passive)	10	9	.90
3.4 adjective (subject-to-subject raising)	120	112	.93
3.5 adjective (subject-to-object raising)	21	18	.86
3.6 noun	101	97	.95
4. *Wh*-word clauses (as subjects or objects)			
4.1 *Wh*-word + *to*-complement	15	14	.93
4.2 *Wh*-word pseudocleft (WH-NP-VP)	79	67	.85
5. Ditransitive clauses			
5.1 Ditransitive (NP-V-NP-NP)	91	82	.91
5.2 Prepositional dative (NP-V-NP-PP)	40	34	.85

I computed agreement between manually annotated results and semi-automatic search results using the measure of recall. In obtaining the recall rate for inter-annotator agreement

between the query results and manual annotation, I followed Lu's (2010, p. 486) procedures, which better address inter-annotator agreement where the annotators determine the unit of coding. The metric was calculated as follows (Hripcsak & Rothschild, 2005; Lu, 2010):

$$\text{Recall} = \frac{\text{Number of Feature X appearing in both R1 and R2}}{\text{Number of Feature X in R1}}$$

In this equation, R1 refers to the list of manually coded items and R2 refers to the list of query results. Recall rate thus shows how many of the desired results (Number of Feature X in R1, annotated as target feature use by the coder) a query was able to retrieve. The recall rate ranged from .85 to 1.00, meaning that the automatic search is estimated to recall 85% to 100% of the desired results.

Accuracy coding. Two coders coded 120 texts each (44 of which were coded by both raters), examining a total of 196 texts for grammatical accuracy. Both coders were native speakers of English with more than 10 years of experiences teaching English as a second language. One coder held a Master's degree in Teaching English as a Second Language (TESOL) and the other was enrolled in a degree program in TESOL at the time of this study. I met with each coder to introduce and trial the coding scheme, which generally followed the guidelines developed by Polio and Shea (2014). In this training session, the coders studied the guidelines and practiced coding with five samples drawn from each possible score level. Next, each coder rated the same set of 20 texts and participated in a follow-up discussion session where the they shared details of their coding decisions to identify discrepantly coded items, determined what constituted an error for coding purposes, and reached agreement on how these items should be coded. Another meeting was held after coding one third of the remaining texts to resolve any issues that arose during independent coding. The revised coding guidelines used in this study are provided in Appendix C.

Depending on the coder's interpretation, one instance of grammatical error could be coded differently in error types and numbers. For example, for the clause "he not tell me nothing about him life," there are at least four possible ways to address the errors: "he tells me nothing about himself," "he tells me nothing about his life," "he doesn't tell me anything about himself," and "he doesn't tell me anything about his life." One coder marked "not" as a verb phrase problem and "him" as the wrong pronoun, presumably suggesting the clause could be revised to read, "he tells me nothing about his life." The other coder marked "not" as a negation problem, "nothing" as a lexical problem, and "him" as the wrong pronoun, which would result in the revised clause, "he doesn't tell me anything about himself." Both coders identified an issue with this clause and their respective identifications of the error were acceptable; however, the resulting error types and error numbers showed variation due to these different interpretations. For this reason, I counted the number of error-free clauses and determined the overall ratio of error-free clauses to total number of clauses in each learner text.

To consistently identify clauses in the texts, I manually divided all 196 texts into clauses using Lu's (2010) definition of clause: a structure with a subject and a finite verb (as also defined by Hunt, 1965; Polio, 1997). I also established some protocols to deal with the difficulties in identifying clauses in learner language: (a) consider a clause with a resumptive pronoun (e.g., "…buy something expensive that *it* is worth waiting for") as one clause; (b) consider such structures clauses even when a subject is missing in a required position (e.g., "If it is something that costs little, *[it]* will not be a thing …"), and (c) consider a string of words that the coders noted to be incomprehensible with unclear structure as just one clause (e.g., "It is one good reason to he/she can is into the fashion and for him/her thinking future in fashion") even if noun phrases that may have been intended as subjects are included. I cross-examined the number of

clauses identified according to this protocol against the output from web software that uses a dependency parser (Lu, 2010) to automatically identify the number of clauses in texts in order to red flag any potential issues with my manual identification of clauses.

As described, to capture potential differences in accuracy across score levels, I used the error-free clause ratio (EFCR) as the measure of accuracy. This ratio is computed by counting the number of error-free clauses and dividing it by the total number of clauses in the target text. The rationale for choosing the EFCR as accuracy measure over the error-free T-unit ratio was that the smaller clausal unit allows the retention of more information. For example, an essay with 10 T-units, 5 of which have one error in each, would have an accuracy score of .5 (5 divided by 10). If the 10 T-units could be further broken down into 20 clauses, the accuracy as measured by EFCR would be .75 (15 divided by 20). In fact, Evans et al. (2014) found that the EFCR had more discriminatory power than the error-free T-unit ratio. Finally, intercoder reliability of the EFCR was computed by comparing the binary coding of error-free clauses. The independent coding of 44 essays showed an 86.1% percent agreement, Pearson correlation $r = .711$ (95% CIs [.671, .747], $p < .001$).

2.3 Data Analysis

The use of 14 target grammatical features was investigated using various methods. Occurrences of each feature are presented in both raw and relative frequencies (per 100 words). Feature occurrences were also broken down by score level. The distribution of feature occurrences was determined by examining how many texts included each feature. The total number of different types of features used per text served as an indicator of syntactic variety.

More specifically, Table 2.9 summarizes the measures described and statistically analyzed in this study. I first examine the characteristics of each target feature by looking at its

raw and relative frequencies (per 100 words) across the five score levels. I performed Kruskal-Wallis tests followed up by Dunn post-hoc tests to reveal any significant differences that may exist between the score levels. Next, I investigated the distribution of the target features in the CELC corpus by analyzing how many texts included each target feature. After examining the overall target feature frequencies and distributions of each feature, I explored the characteristics of each text by computing the raw and relative frequencies of all target features per text and the number of different types of features used per text. These measures served as independent variables in multiple linear regressions to predict the score level of the text. Assumptions such as linearity, multicollinearity, normal distribution of the residuals, and homoscedasticity were examined after performing each multiple linear regression.

Table 2.9

Measures and Statistical Analysis of the Full Corpus (N = 560)

	Analysis	**Feature occurrences and frequencies**	**Distribution**
Characteristics of each feature across score levels (k = 14)	Descriptive	- Raw frequencies - Relative frequencies	- Number of texts - Proportion of texts
	Significance test	Kruskal-Wallis test (non-parametric) - DV: relative frequencies - IV: score level Post hoc: Dunn test	
Feature use pattern for each text within each score level (n = 560)	Descriptive	- Raw number of target features - Relative frequencies of target features - Number of different types of features used	- Contingency table of occurrences
	Inferential	Regression analysis - DV: score level - IVs: raw frequencies of target features, relative frequencies of target features, number of different types of features used	- Principal components analysis

Using more advanced statistical analysis, I aimed to discover the co-occurring linguistic patterns in the data and how such patterns are associated with the given scores. While a number of options are available to identify which features tend to occur together (e.g., cluster analysis, discriminant function analysis, exploratory factor analysis), I considered principal components analysis (PCA) to be a statistical method useful for the purpose of this study due to the fact that PCA not only provides a summary of which grammatical features pattern together but also computes component scores for each text. The resulting data can then be fed into another regression model to examine to what extent the components predict grammar score level. Many previous studies have adopted multidimensional analysis, which utilize exploratory factor analyses (e.g., Biber, 1988; Egbert & Biber, 2018; Staples, LaFlair, & Egbert, 2017). For these studies, underlying factors, such as different modalities (i.e., speaking and writing) and task types (e.g., Biber et al., 2016) or registers (e.g., LaFlair & Staples, 2017), that "cause" patterns of linguistic feature use were posited. In this study, I assume no underlying construct but rather attempt to discover correlated grammatical features by analyzing all the variances in the observed data. For this purpose, I used PCA, which is also recommended as a psychometrically more advisable procedure (Field, Miles, & Field, 2012) and for large studies (Biber & Egbert, 2016).

Data analyses performed for the data subset are described in Table 2.10. As with the analysis of the full corpus, I examined the same feature frequencies and number of type measures per text. This time, however, the EFCR was included as one of the independent variables in multiple linear regressions to predict the score level.

Table 2.10

Measures and Statistical Analysis for Subset (n = 196)

	Analysis	Feature occurrences and frequencies	Errors
Feature use pattern for each text within each score level (n = 196)	Descriptive	- Raw number of target features - Relative frequencies of target features - Number of different types of features used	- Error-free clause ratio (EFCR)
	Inferential	Regression analysis - DV: score level - IVs: total number of target features, relative frequencies of target features, number of different types of features used, EFCR	

I used multiple linear regression to predict the grammar score levels that ranged from 1 to 5. Binary logistic regression was also used to examine how well a set of predictors differentiate the two adjacent score levels. Logistic regression estimates the probability of a given binary outcome (in this case, the two levels being compared) taking place as a function of the predictors (Gries, 2013; Tabachnick & Fidell, 2013). This allowed a break-down of the rating scale to investigate to what extent the predictor variables could distinguish between adjacent scores.

CHAPTER 3: RESULTS

In this chapter, I describe the use of the 14 target grammatical features at the individual feature level and move on to patterns of use (i.e., frequencies and types) by each text. Next, I identify co-occurring features and investigate to what extent these patterns associate with score level. I also use a subset of data coded for accuracy in order to examine the effect of linguistic accuracy on score level. Finally, I examine how well the use of grammatical features and accuracy distinguish between adjacent score levels.

3.1 Target Grammatical Feature Use

This study examined the use of 14 specific grammatical structures. To illustrate how these features are attested in the CELC corpus, I present (a) the frequencies of the features, (b) the distribution of the features, and (c) patterns of feature co-occurrence.

3.1.1 Frequencies of the target grammatical features

The frequencies of investigated features aggregated by score levels are shown in Table 3.1, with the mean relative frequencies per 1000 indicated in parenthesis. To account for the varying size of the subcorpora and each text, I discuss the results using relative frequencies. Relative clause (F1.1) and *that*-complement clauses controlled by verb (F2.1) occurred most frequently, followed by *to*-complement clauses controlled by verb (F3.1). The relative frequencies of relative clauses showed some differences at the lower score levels, with the frequency being higher at score level 3 (1.31) than at score level 2 (1.02), which was higher than the frequency at score level 1 (0.71). Non-parametric Kruskal-Wallis rank sum test and a post hoc Dunn test indicated that significant differences existed among the scoring groups ($\chi^2_{(4)}$ = 41.20, $p < .001$) for this feature, specifically between score levels 1 and 2, between score levels and 2 and 3. The frequencies of *that*-complement clauses with verb were high overall, ranging between 0.96 and 1.05, but did not show significant differences across score levels. The frequency of *to*-clauses controlled by verb showed some difference between score levels 1 (0.48) and 2 (0.70), as confirmed by a Kruskal-Wallis test ($\chi^2_{(4)}$ = 12.10, p = .017) and a post hoc test (z = -2.535, p = .034, Benjamini & Hochberg corrected).

Table 3.1

Frequencies of Target Grammatical Features by Each Score Level

(mean relative frequency per 100 words)

Feature description	Score 1 (n = 54)	Score 2 (n = 137)	Score 3 (n = 158)	Score 4 (n = 146)	Score 5 (n = 65)	Mean
1. Post-nominal modification with						
1.1 Relative clause	68 (0.71)	282 (1.02)*	493 (1.31)*	566 (1.50)	264 (1.49)	1673 (1.24)
2. *That*-complement clauses controlled by						
2.1 verb	89 (0.98)	262 (0.96)	359 (0.97)	398 (1.04)	193 (1.05)	1301 (0.99)
2.2 adjective	4 (0.04)	10 (0.04)	13 (0.04)	7 (0.02)	9 (0.06)	43 (0.04)
2.3 noun	4 (0.06)	3 (0.01)	8 (0.02)	6 (0.01)	5 (0.03)	26 (0.02)
3. *To*-complement clauses controlled by						
3.1 verb	44 (0.48)	196 (0.70)*	285 (0.77)	296 (0.76)	116 (0.66)	937 (0.71)
3.2 verb (subject-to-object raising)	12 (0.13)	65 (0.23)	124 (0.33)	158 (0.41)	72 (0.43)	431 (0.32)
3.3 verb (passive)	0	0	3 (0.01)	6 (0.02)	3 (0.02)	12 (0.01)
3.4 adjective	5 (0.06)	35 (0.12)	49 (0.13)	35 (0.10)	22 (0.12)	146 (0.11)
3.5 adjective (subject-to-object raising)	0	8 (0.03)	19 (0.05)	19 (0.04)	10 (0.06)	56 (0.04)
3.6 noun	10 (0.09)	42 (0.16)	90 (0.24)*	121 (0.31)	51 (0.28)	314 (0.23)
4. *Wh*-word clauses (as subjects or objects)						
4.1 WH + *to*-complement	0	7 (0.02)	13 (0.03)	17 (0.04)	5 (0.02)	42 (0.03)
4.2 WH-NP-VP	5 (0.06)	32 (0.10)	82 (0.22)*	95 (0.26)	43 (0.24)	257 (0.19)
5. Ditransitive clauses						
5.1 Ditransitive (NP-V-NP-NP)	14 (0.13)	37 (0.14)	67 (0.19)	56 (0.15)	38 (0.21)	212 (0.16)
5.2 Prepositional dative (NP-V-NP-PP)	8 (0.08)	16 (0.06)	32 (0.08)	34 (0.09)	11 (0.06)	101 (0.08)

Note. * Indicates statistically significant difference from the lower adjacent score.

Among the less frequently used features, the frequency of *to*-clauses controlled by noun (F3.6; e.g., They have *food to eat*.) showed some difference between score levels 2 and 3 (1.6 and 2.4 per 1000 words, respectively), as confirmed by a Kruskal-Wallis test ($\chi^2_{(4)} = 26.58$, $p = <.001$) and a post hoc test ($z = -2.23$, $p = .043$, Benjamini & Hochberg corrected). Similarly, the relative frequencies of *wh*-word pseudocleft clause (F4.2; e.g., *What I want to say* is this.)

indicated a significant difference between score levels 2 and 3 (1.0 and 2.2 per 1000 words, respectively): Kruskal-Wallis test, $\chi^2_{(4)} = 33.10$, $p = <.001$; Dunn test, $z = -3.25$, $p = .002$, Benjamini & Hochberg corrected. Full results of significant tests are reported in Table D1, Appendix D.

Features that showed significant differences between score levels are visually summarized in Figure 3.1. The contour of each violin plot shows the distribution of the data. For instance, in the F1.1 panel, data points are highly concentrated at the frequency of 0 at Score 1 (see the left-most plot) even though the mean relative frequency is closer to 1, causing the plot to resemble a triangle with a wide base. The distribution becomes less skewed at the higher scores. Therefore, the plot illustrates that along with higher mean frequencies, the relative frequencies of relative clauses become more evenly distributed around the mean at the higher scores. The three horizontal lines within each violin plot indicate interquartiles: 25%, 50% median, and 75% respectively from bottom to top. The red dot indicates the mean, and error bars indicate bootstrapped 95% confidence intervals.

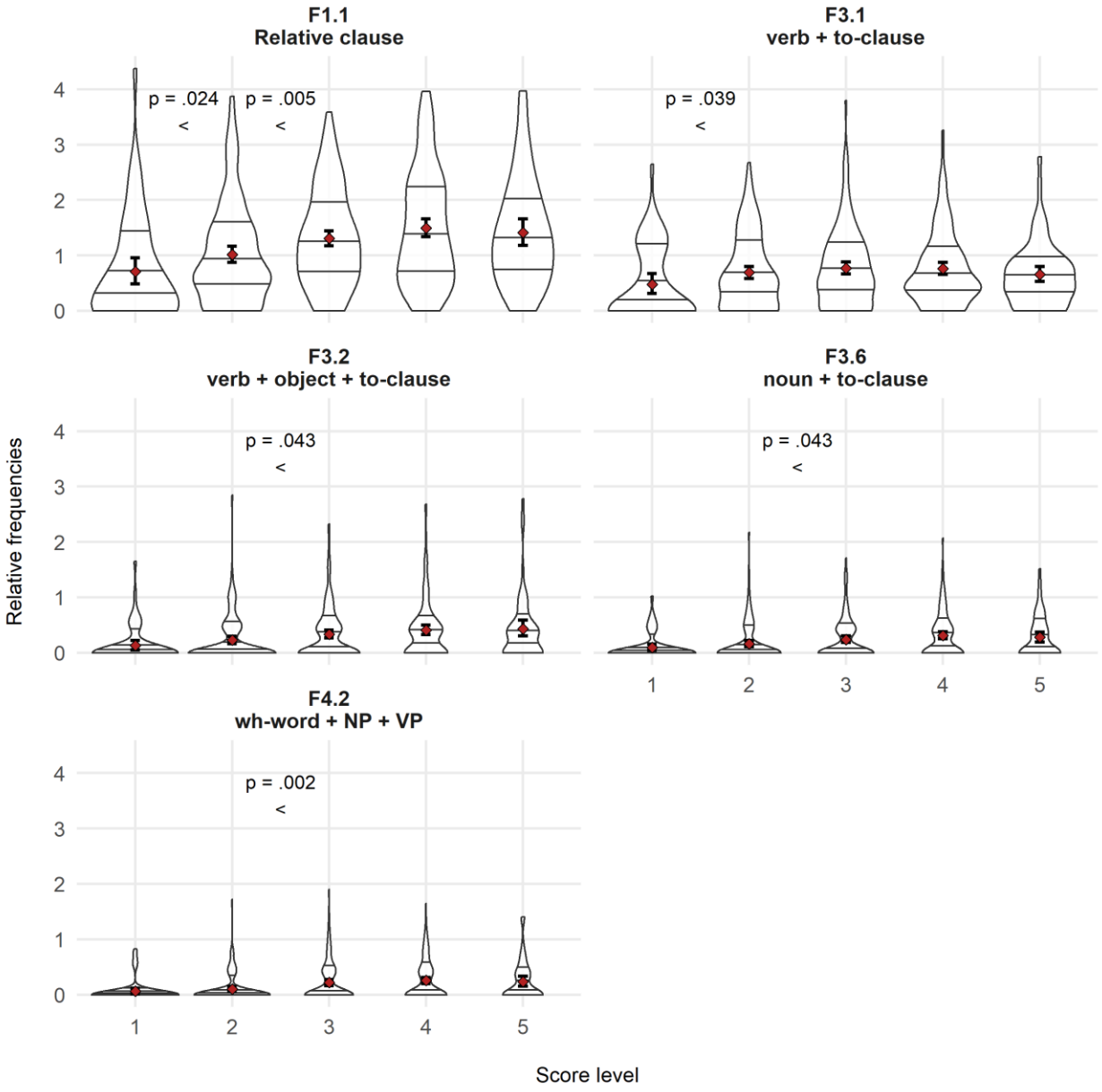

Figure 3.1. *Relative frequencies of grammatical features displaying statistically significant differences across score levels.*

In summary, the frequencies of three features (F1.1, F2.1, and F3.1) were high while the other features appeared relatively infrequently. No features showed a statistically significant, linear increase across score levels; however, six features showed statistically significant differences between sets of adjacent score levels, namely, levels 1 and 2 and/or levels 2 and 3.

3.1.2 Distribution of grammatical features

While the frequencies reported in the previous section depict how often a particular feature appears in texts at a certain level, they do not provide a fuller picture of how widely the features manifest in the corpus, i.e., what proportion of the texts include a particular feature. To examine the distribution of the grammatical features throughout the corpus, I present the number and proportion of texts that include each feature at least once.

As shown in Figure 3.2, the most frequently employed feature was relative clause construction (F1.1), which appeared in 85.7% of the 560 texts. *That*-complements controlled by a verb (F2.1) were used in 84.3% of the texts, and *to*-complements controlled by a verb (F3.1) appeared in 73.4% of the texts. In other words, these three features were used not only frequently, but also widely.

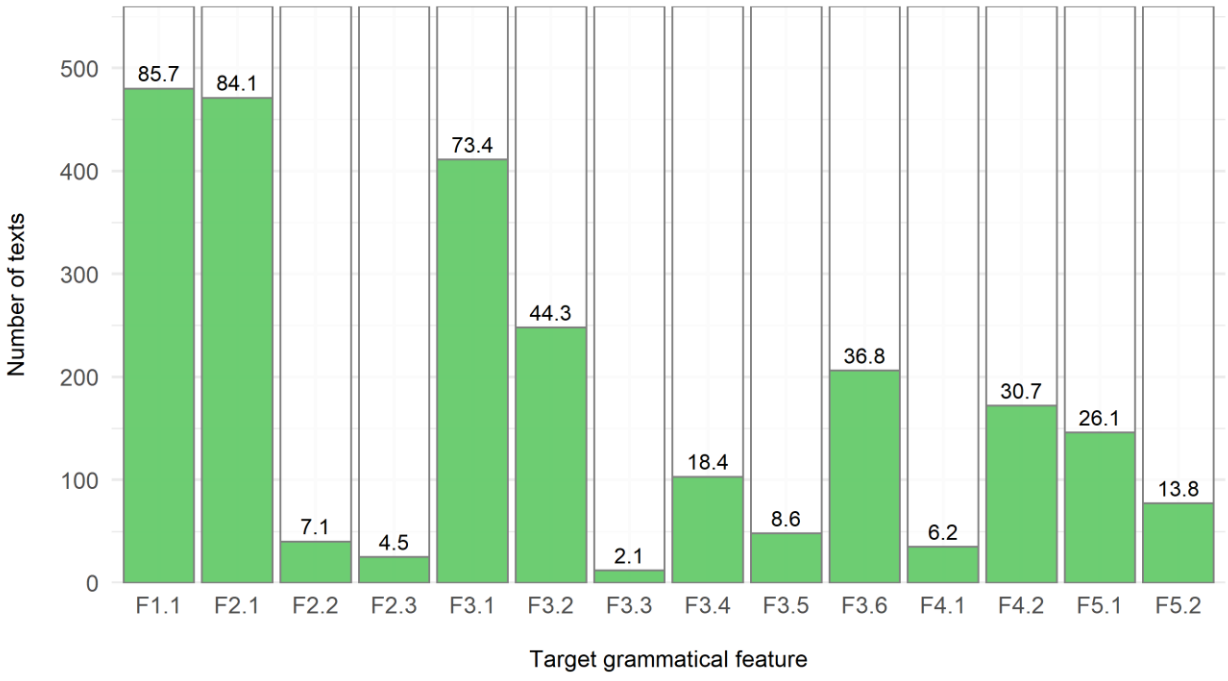

Figure 3.2. *Number and proportion of texts including each target grammatical feature.*

To-complements controlled by verb in raising constructions (F3.2; e.g., "I *found the task to be* much more difficult.") and by noun (F3.6; e.g., "They have *food to eat*.") were relatively

widely used, appearing in 40.4% and 36.8% of the texts, respectively. In contrast to the frequently and widely used *that*-clauses controlled by verb (F2.1), *that*-clauses controlled by adjectives and nouns occurred in a much smaller number of texts, 7.1% and 4.5% of the corpus, respectively. Used in 3.9% of all corpus texts, *to*-complements controlled by passive verb constructions (F3.3) were attested in fewer texts than any other features examined. Table 3.2 breaks down these numbers by score level: The top row in each cell indicates the number and proportion of texts containing the given feature in each score level (reflected in each column); and the bottom row indicates how many times, on average, the given feature appeared in a single text.

Table 3.2

Number and Proportion of Texts at Each Score Level Displaying Each Grammatical Feature

Feature	**Score 1 (n = 54)**	**Score 2 (n = 137)**	**Score 3 (n = 158)**	**Score 4 (n = 146)**	**Score 5 (n = 65)**	**Total**
1. Post-nominal modification with						
1.1 Relative clause	30 (55.6) 1.3	108 (78.8) 2.1	142 (89.9) 3.1	138 (94.5) 3.9	62 (95.4) 4.1	480 (85.7)
2. *That*-complement clauses controlled by						
2.1 verb	39 (72.2) 1.6	107 (78.1) 1.9	135 (85.4) 2.3	131 (89.7) 2.7	59 (90.8) 3.0	471 (84.3)
2.2 adjective	4 (7.4) 0.1	10 (7.3) 0.1	12 (7.6) 0.1	7 (4.8) < 0.1	7 (10.8) 0.1	40 (7.1)
2.3 noun	4 (7.4) 0.1	3 (2.2) < 0.1	8 (5.1) 0.1	5 (3.4) < 0.1	5 (7.7) 0.1	25 (4.5)
3. *To*-complement clauses controlled by						
3.1 verb (subject-to-subject raising)	23 (42.6) 0.8	95 (69.3) 1.4	120 (75.9) 1.8	121 (82.9) 2.0	52 (80.0) 1.8	411 (73.4)
3.2 verb (subject-to-object raising)	10 (18.5) 0.2	43 (31.4) 0.5	72 (45.6) 0.8	85 (58.2) 1.1	38 (58.5) 1.1	248 (40.4)
3.3 verb (passive)	-	-	3 (1.9) < 0.1	6 (4.2) < 0.1	3 (4.6) < 0.1	12 (3.9)

Table 3.2 (cont'd)

3.4 adjective (subject-to-subject raising)	5 (9.3) 0.1	22 (16.1) 0.3	34 (21.5) 0.3	29 (19.9) 0.2	13 (20.0) 0.3	103 (18.4)
3.5 adjective (subject-to-object raising)	-	7 (5.1) 0.1	17 (10.8) 0.1	16 (11.0) 0.1	8 (12.3) 0.2	48 (8.6)
3.6 noun	9 (16.7) 0.2	33 (24.1) 0.3	61 (38.6) 0.6	73 (50.0) 0.8	30 (46.2) 0.8	206 (36.8)
4. Wh-word clauses (as subjects or objects)						
4.1 WH + *to*-complement	0	7 (5.1) 0.1	13 (8.2) 0.1	10 (6.8) 0.1	5 (7.7) 0.1	35 (6.2)
4.2 WH-NP-VP	5 (9.3) 0.1	24 (17.5) 0.2	55 (34.8) 0.5	61 (41.8) 0.7	27 (41.5) 0.7	172 (30.7)
5. Ditransitive clauses						
5.1 Ditransitive (NP-V-NP-NP)	7 (13.0) 0.3	28 (20.4) 0.3	48 (30.4) 0.4	39 (26.7) 0.4	24 (36.9) 0.6	146 (26.1)
5.2 Prepositional dative (NP-V-NP-PP)	4 (7.4) 0.1	14 (10.2) 0.1	21 (13.3) 0.2	29 (19.9) 0.2	9 (13.8) 0.2	77 (13.8)

The numbers show that, overall, the proportion of texts including a given feature tended to be larger at higher score levels. The average number of target grammatical features in a text is visualized in Figure 3.3.

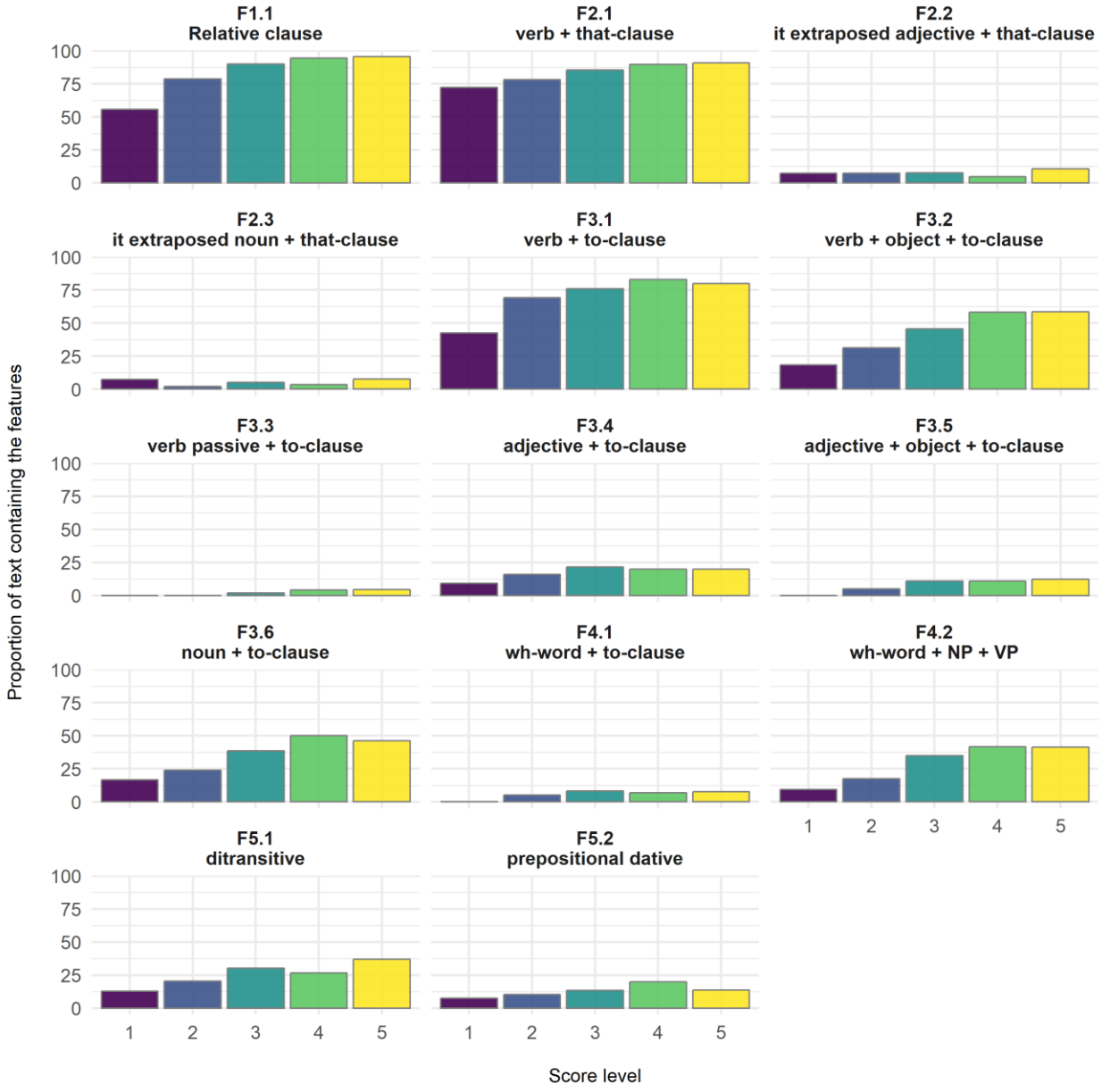

Figure 3.3. *Proportion of texts at each score level shown by target feature.*

The number of times a feature appeared in a text tended to increase across score levels, more notably through the lower score levels (see Figure 3.4). Note that this frequency is not normalized in relation to the length of the texts but provides insight into how many features an average text included at each score level. The frequency of features per text will be statistically examined in relation to the score level in Section 3.2.2.

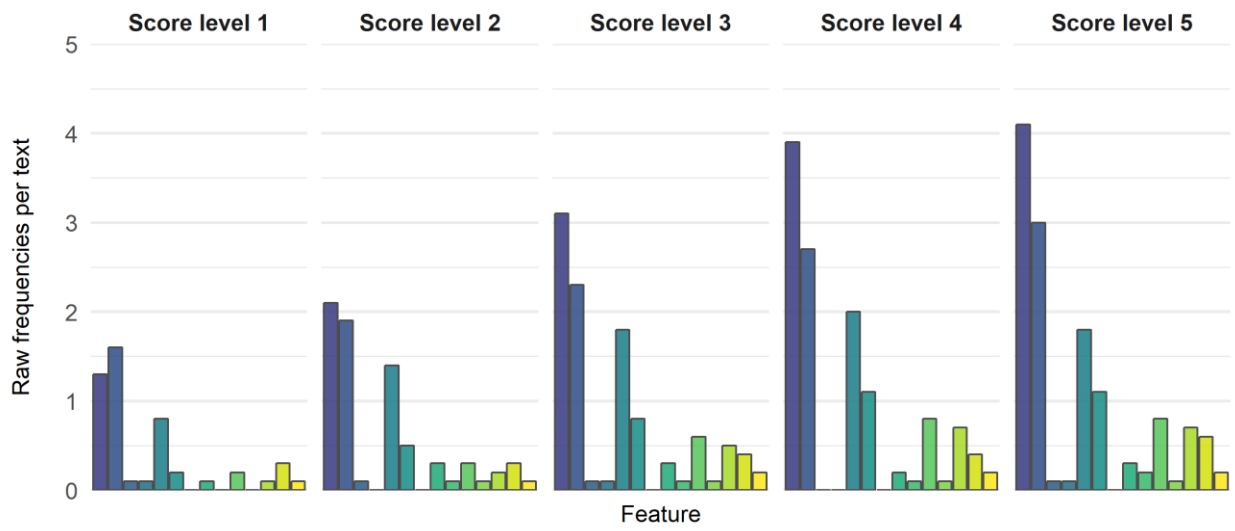

Figure 3.4. *Average number of target grammatical features per text.*

To summarize, the frequencies and distributions of features corresponded to one another, with more frequently occurring grammatical features also appearing in a larger proportion of the texts in the corpus. It is to be expected that, in general, some grammatical constructions are used more often than others. The descriptive statistics on the distribution of features and average number of features used per text suggested a tendency that the features were more frequently and widely used in texts at higher score levels than in those at lower levels. Moving on from use of each individual target grammatical feature, I turn now to the combined use of features.

3.1.3 Patterns of co-occurrence of target grammatical features

In order to investigate which of the 14 features pattern together, I performed a PCA. The outcomes of a PCA are the *components* (aggregates of correlated variables that summarize the given data [Tabachinick & Fidell, 2013]), which are discovered through analyzing all variance in the data. For this study, I used the dichotomously coded data for the 14 target features in 560 texts. As PCA is purely a mathematical procedure, it is important that the results make sense to the researcher (Tabachinick & Fidell, 2013). I opted for the occurrence/non-occurrence (i.e., whether a feature was used in a text or not) binary coding instead of using frequencies of the

features, because the results were uninterpretable with the frequency data. Using IBM SPSS Statistics 25, I extracted six components with Promax (oblique) rotation, which accounted for about 53% of the total variance in the data.

Table 3.3

Structure Matrix of Principal Components Analysis (large loadings highlighted in grey)

	Component					
Feature	**1**	**2**	**3**	**4**	**5**	**6**
F1.1 Relative clause	.593	.085	.152	.112	-.009	.172
F2.1 verb + *that*-clause	.094	.613	.111	.131	-.173	.070
F2.2 *it* extraposed adjective + *that*-clause	-.464	.149	.473	.253	.008	-.475
F2.3 *it* extraposed noun + *that*-clause	.115	-.015	.028	.094	.741	-.051
F3.1 verb + *to*-clause	.351	.244	.109	.207	.092	.341
F3.2 verb + object + *to*-clause	.181	.321	-.187	.675	-.174	.174
F3.3 verb passive + *to*-clause	.026	-.174	.162	.615	.225	-.044
F3.4 adjective + *to*-clause	.064	-.026	.631	-.269	.117	-.117
F3.5 adjective + object + *to*-clause	.390	.213	-.146	.223	-.474	-.164
F3.6 noun + *to*-clause	.642	.018	-.050	.039	.123	-.177
F4.1 *wh*-word + *to*-clause	.054	.035	.618	.168	-.018	.118
F4.2 *wh*-word + NP + VP	.161	.568	-.057	.049	.503	.215
F5.1 ditransitive	-.054	.053	-.004	.035	-.002	.700
F5.2 prepositional dative	.195	-.512	.235	.317	-.095	.418

I summarize what each component encapsulates, by interpreting the larger loadings (> | ±.5|, highlighted in gray).

- Component 1: Use of relative clauses (.593), *to*-complement clauses controlled by a noun (.642)
    - This component summarizes the use of complex noun phrases, such as nouns modified by relative clauses and nouns modified by to-complement clauses.

- Component 2: Use of verb *that*-complements (.613), *wh*-word clauses (.568), and less use of dative constructions (-.512)
    - This component summarizes embedded finite clauses and less complex clause using prepositional phrases.

- Component 3: Use of *to*-complements with *it* extraposed adjective (.631), *to*-complements with *wh*-word (.618)
    - This component summarizes the use of non-finite verb phrases, the use of to-complement with it extraposition and with wh-words.

- Component 4: Use of verb *to*-complements with raising construction (.675), *to*-complements with passives (.615)
    - This component summarizes the use of relatively complex to-complement controlled by verbs, with an object as the semantic argument of the to-infinitive verb, or with a passive verb form.

- Component 5: Use of *that*-complements controlled by noun (.741) and *wh*-word pseudocleft (.503)
    - This component summarizes use of relatively complex noun phrases with finite clauses.

- Component 6: Use of ditransitives (.719)
    - This component mainly summarizes the use of ditransitive constructions with both a direct and an indirect object.

After the six components were extracted by performing a PCA, component scores were computed. A component score utilizes weights (i.e., coefficients computed from factor/component loadings) to calculate a sum of weights multiplied by the given (standardized) data. In other words, component scores computed for this data served as a summary of grammatical feature use for each text. Figure 3.5 illustrates the distribution of the six component scores across score levels.

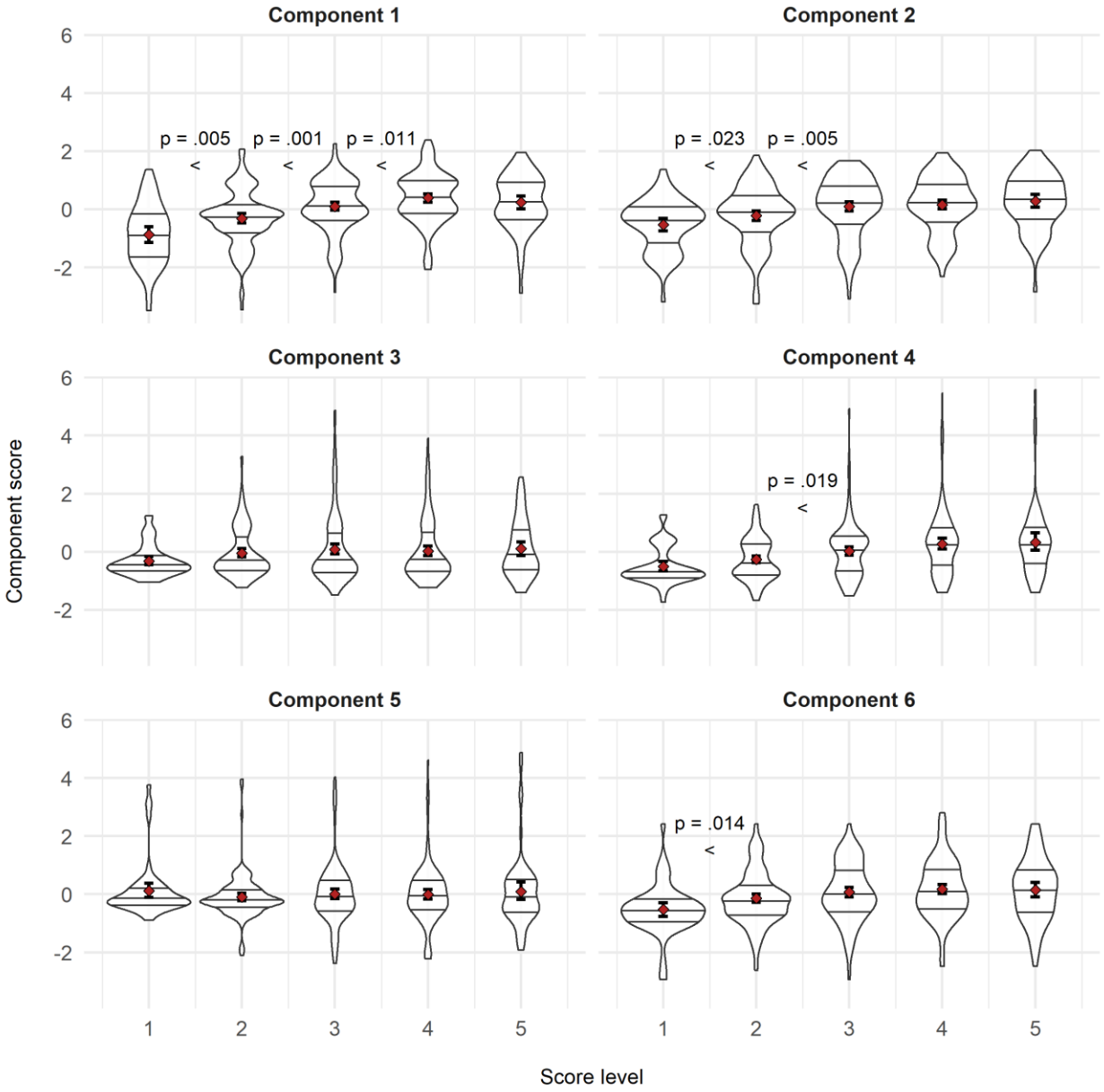

Figure 3.5. *Component scores across score levels.*

To examine the differences in component scores across score levels, I performed a Kruskal-Wallis test followed by a Dunn post hoc (Benjamini & Hochberg corrected). The post hoc results between adjacent levels are indicated in Figure 3.5 and the full results are reported in Table D2 of Appendix D. As seen, significant differences were detected in the use of complex noun phrases (Component 1) among all score levels except for score level 5. The use of embedded finite clauses (Component 2) showed significant differences between score levels 1

and 2 and between score levels 2 and 3. The use of complex *to*-complements (Component 4) differed between score levels 2 and 3, and the use of ditransitive clauses (Component 6) between score levels 1 and 2.

To examine how these patterns of grammatical feature co-occurrence help predict score levels, I performed a multiple linear regression with the component scores as independent variables.

Table 3.4

Multiple Linear Regression with Component Scores as Predictors

	Predictor variables	***B***	***Std error***	***Beta***	***t-value***	***p-value***	R^2	ΔR^2
Model 1	(Intercept)	3.06	0.04		70.53	< .001	.232	.224
	Component 1	0.35	0.05	0.30	7.68	< .001		
	Component 2	0.24	0.04	0.20	5.35	< .001		
	Component 3	0.14	0.05	0.12	3.24	.001		
	Component 4	0.22	0.05	0.19	4.99	< .001		
	Component 5	0.04	0.04	0.03	0.91	.363		
	Component 6	0.13	0.05	0.11	2.87	.004		
Model 2 (final)	(Intercept)	3.06	0.43		70.54	< .001	.231	.224
	Component 1	0.35	0.05	0.30	7.67	< .001		
	Component 2	0.24	0.04	0.20	5.37	< .001		
	Component 3	0.15	0.04	0.13	3.35	.001		
	Component 4	0.22	0.04	0.19	4.92	< .001		
	Component 6	0.13	0.05	0.11	2.91	.004		

Except for Component 5, all components were significantly predictive of score level. The final model (Model 2) was significant, explaining about 23.1% of the variance in grammar subscore, $F_{(5, 554)} = 147.34$, $p < .001$, $R^2 = .231$, $\Delta R^2 = .224$. The regression analysis results indicated that use of grammatical features positively correlates with assigned grammar score, particularly: use of complex noun phrases (Component 1), use of embedded finite clauses (Component 2), use of non-finite verb phrases such as *to*-complement with *it* extraposition and with *wh*-words (Component 3), use of more complex *to*-complements (Component 4), and

ditransitive clauses (Component 6).

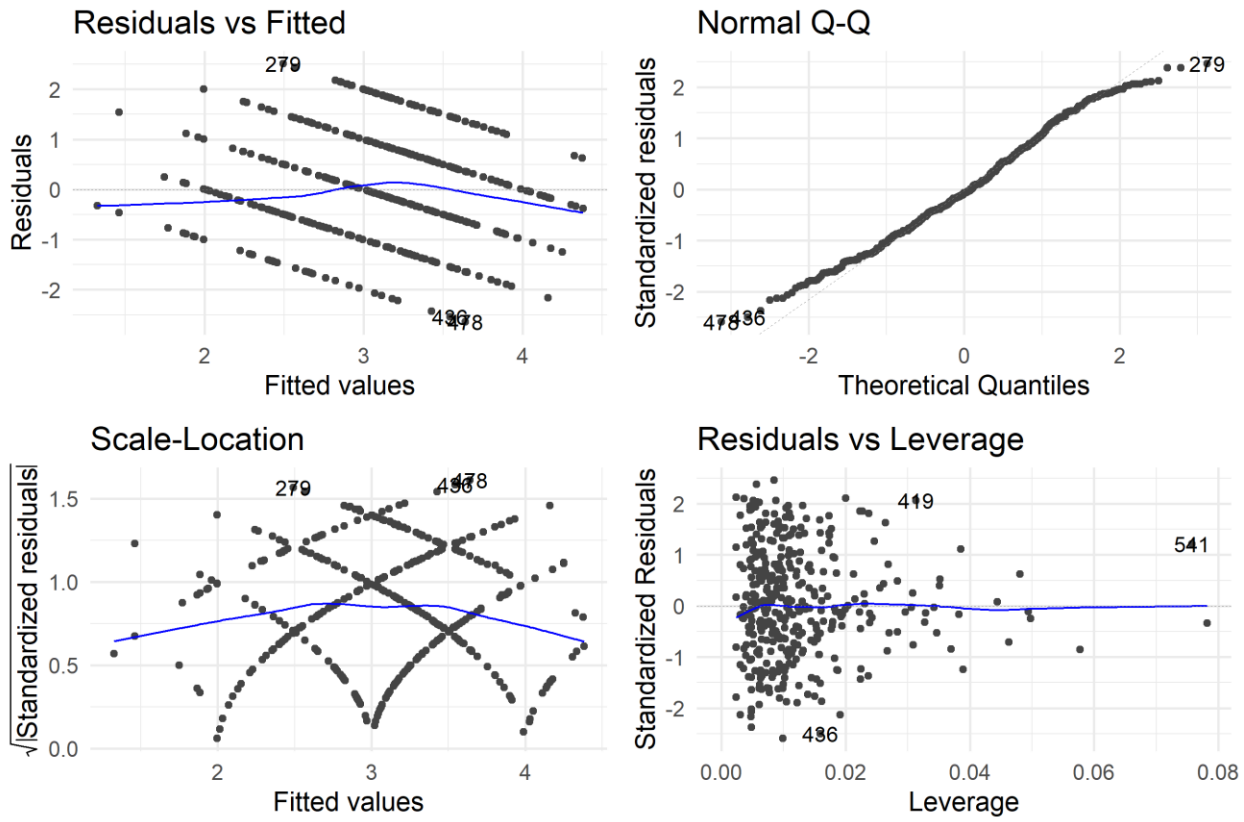

Figure 3.6. *Diagnostic plots for the multiple linear regression model.*

Figure 3.6 presents four plots for visually inspecting the assumption of linearity, normal distribution of residuals, homogeneity of variance, and absence of influential cases. The top 3 most extreme points are automatically labeled in the plots the labeled points are not implied to be in violation of the assumptions. The top-left plot shows the spread of residuals. If the assumption of linearity is met, the line is approximately horizontal at zero. The line in this case is mostly flat, suggesting a linear relationship between the predictor variables and the outcome variable. Homogeneity of variance is checked by the scale-location plot (bottom-left panel), the relatively flat line in which indicates homoscedasticity. Normality of residuals is seen in the Q-Q plot (top-right panel). Observations with standardized residuals greater than 3 in absolute value are possible outliers (James, Witten, Hastie, & Tibshirani, 2014), but no data points exceed ± 3.

Lastly, the plot examining influential data points (bottom-right panel) does not display Cook's distance lines, meaning that all points are within an acceptable range of Cook's distance and there are no influential data points. In addition, the model fit was assessed to detect potential issues of overfitting by influential data points by bootstrapping the samples 5,000 times. The result yielded only a minor difference from the original model (corrected R^2 = .218), indicating that there was no overfitting of the model caused by influential observations.

3.2 The Relationship Between Feature Use and Score Levels

In this section, I explore the use of grammatical features in relation to assgined grammar subscore. For this purpose, I first report descriptive statistics for the number of different types of target grammatical features and their frequencies, and investigate how well the measures predict score level through regression analyses.

3.2.1 Descriptive statistics

Table 3.5 illustrates (a) the aggregated mean frequency of the 14 features, (b) the mean relative frequency (per 100 words) of the features per text, (c) the mean number of different features used (henceforth referred to as feature types), and (d) the mean number of words per text. The table is accompanied by Figure 3.7 for visual interpretation of the data.

Table 3.5

Target Feature-Related Measures and Number of Words by Score Level

	Mean raw feature frequency (SD)	**Mean relative frequency (SD)**	**Mean number of feature type (SD)**	**Mean number of words (SD)**
Score level 1	4.8 (3.14)	2.78 (1.71)	2.57 (1.47)	173 (41.2)
Score level 2	7.2 (3.83)	3.54 (1.53)	3.62 (1.48)	200 (42.6)
Score level 3	10.3 (4.18)	4.34 (1.56)	4.66 (1.44)	236 (48.7)
Score level 4	12.3 (4.59)	4.72 (1.46)	5.14 (1.48)	261 (53.7)
Score level 5	12.9 (4.13)	4.62 (1.46)	5.26 (1.25)	284 (52.8)

Both feature frequencies and number of feature types increased as the score level advanced from 1 through 4. The differences were minimal between score levels 4 and 5. One finding of interest was that the mean number of feature types (third column of Table 3.5) used at score level 1 was 2.57. At score level 2, the number of feature types was 3.62, an increase representing learner use of approximately one additional feature type compared to the adjacent lower level. Likewise, another additional feature type was used at score level 3, as texts at this level featured about 4.66 feature types on average. These numbers indicate that feature frequency and number of feature types can be useful in differentiating the lower score levels, but not so much the higher score levels. Figure 3.7 visualizes these four measures, displaying the distribution (data points and violin plots), mean for each score level (indicated in diamonds), and 95% confidence intervals on the mean (indicated in error-bars).

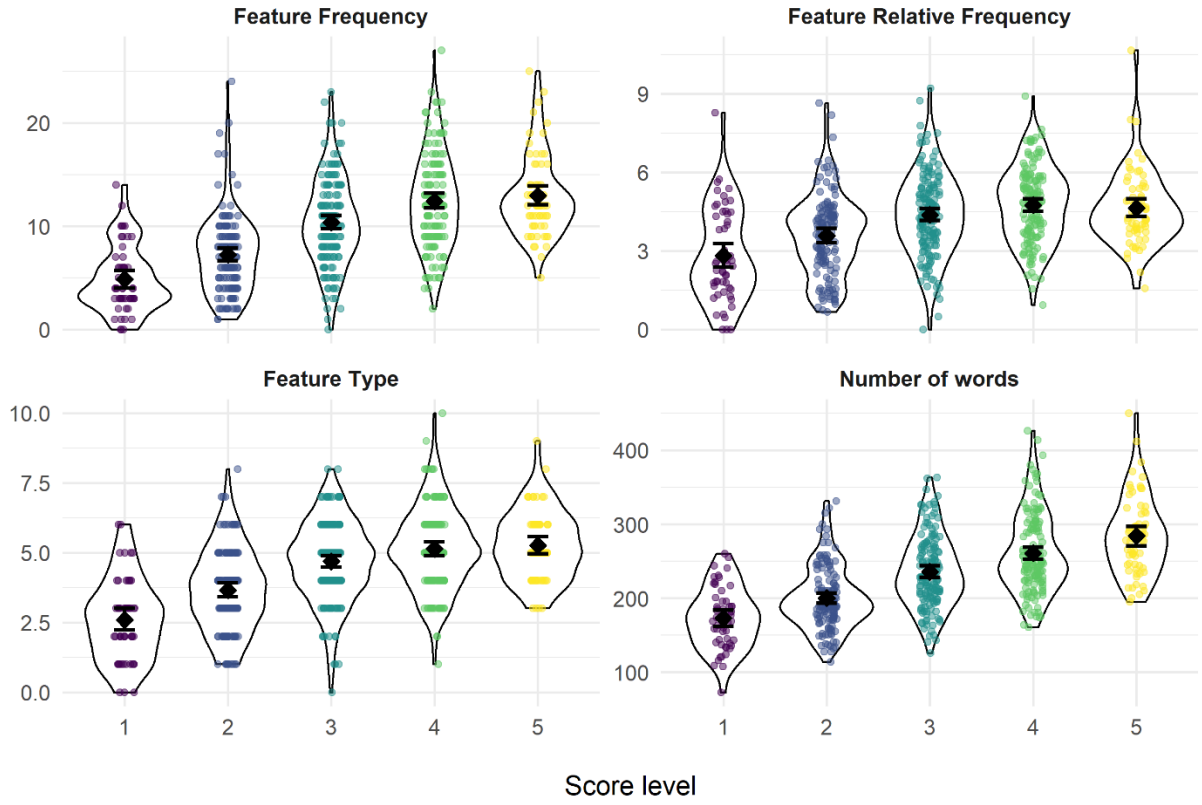

Figure 3.7. *Target feature-related measures by score level.*

3.2.2 Regression analysis for predicting grammar subscore

To statistically examine how predictive grammatical feature use was of grammar subscore, I performed multiple linear regressions. The regression model with component scores provided limited information about the use of the features, as the component scores do not reflect the frequencies of feature occurrence. I performed a regression analysis with feature frequency and number of feature types (indicating syntactic range) as predictor variables. Firstly, I examined the relationship among the three measures, as shown in Figure 3.8.

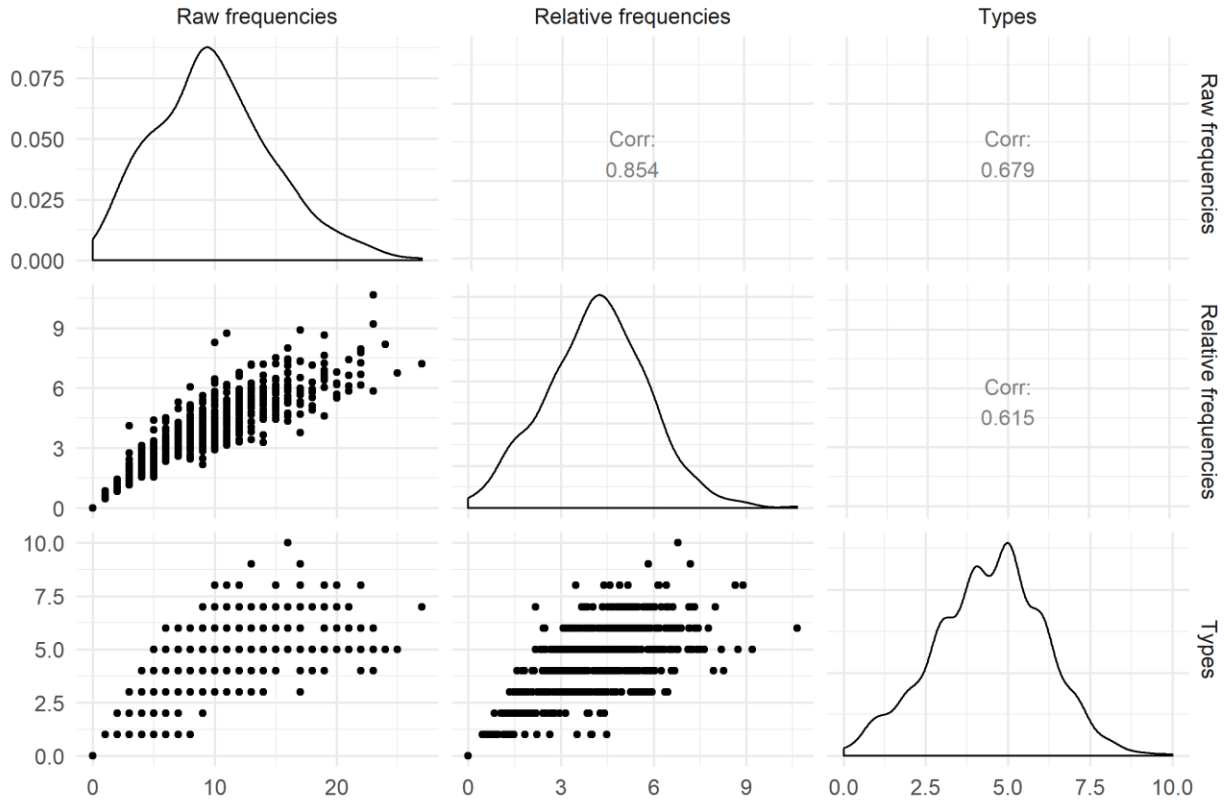

Figure 3.8. *Correlation and scatterplots of the possible predictor variables.*

The raw frequencies and relative frequencies, not surprisingly, highly, linearly correlated (r = .854). The number of different feature types used in each text and the two frequency measures were moderately correlated (r = .679 for raw frequency and r = .615 for relative frequency). For the following regression analysis, multicollinearity was checked at every step by computing the variance inflation factor (VIF). Variables in all final models did not exceed a VIF value of 2, which is considered to be within an acceptable range (Levshina, 2015).

Table 3.6

Multiple Linear Regression Modeling Predicting Score Level

	Variables	*B*	*Std error*	*Beta*	*t-value*	*p-value*	R^2	ΔR^2
Model 1	(Intercept)	1.58	0.12		12.82	< .001	.228	.227
	Feature type	0.36	0.03	0.48	12.84	< .001		
Model 2	(Intercept)	1.49	0.13		11.34	< .001	.232	.230
	Feature relative frequency	0.06	0.03	0.08	1.75	.081		
	Feature type	0.30	0.03	0.43	9.07	< .001		
Model 3 (final)	(Intercept)	1.48	0.12		12.58	< .001	.303	.301
	Feature raw frequency	0.09	0.01	0.37	7.76	< .001		
	Feature type	0.16	0.03	0.22	4.64	< .001		

Table 3.6 describes regression models for predicting the score level. I started with the number of different feature types as a single predictor in Model 1, which was statistically significant in predicting the score level. In Model 2, the addition of relative feature frequency had marginal impact on the model, as the variable was not found to be a statistically significant predictor of the score level. However, the raw frequency (total number of any features used in each text) served as a strong predictor of the grammar subscore in Model 3, substantially improving the fit of Model 1. Although relative feature frequency makes possible objective comparisons of texts of different lengths, given the fact that the grammar subscores are assigned by human raters who read each text in its entirety, it may be more ecologically valid to instead consider the raw frequencies of the target features.

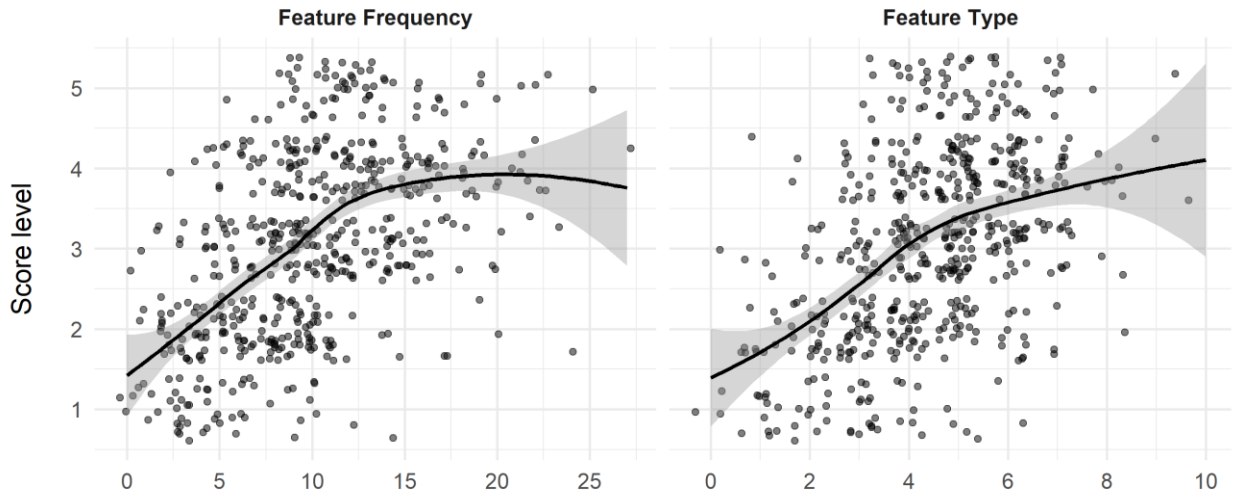

Figure 3.9. *Scatterplots of score levels, feature frequency and number of feature types.*

The final model included number of feature types and feature frequency as two significant predictors of grammar subscore ($F_{(2, 557)} = 121.30$, $p < .001$, $R^2 = .303$, $\Delta R^2 = .301$). In other words, the number of different types of the grammatical features used and the frequency of these features explained about 30.3% of the variance in score level.

Although not presented here, I performed a multiple linear regression analysis that also included the component scores as predictor variables for the grammar subscore. The component scores turned out not to be significant in predicting the grammar score when the number of types of features and frequencies were included together in the same model. This may suggest that measures representing overall diverse use of grammatical features may be more effective in predicting the score level than measures reflecting uses of certain grammatical constructions.

The assumptions for multiple linear regression were visually examined with plots shown in Figure 3.10. The assumptions of random errors and homoscedasticity were inspected by plotting the standardized residual against the predicted value (top-left panel). The line slightly moves away from the zero value as the fitted value increases, suggesting the possibility of a non-linear relationship between the predictor variables and outcome variable.

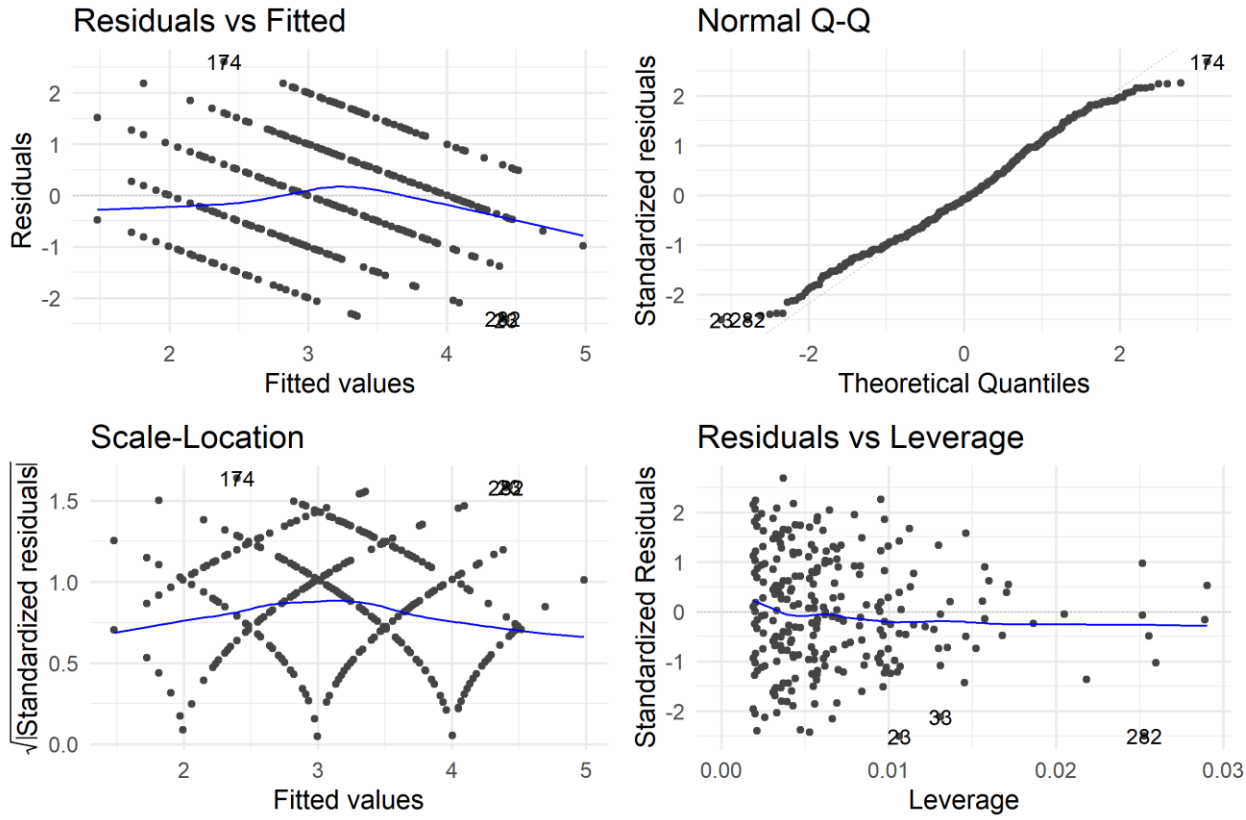

Figure 3.10. *Residual plots for visual examination of linear regression assumptions.*

Given this possibility of non-linearity, I performed a polynomial regression to model potential non-linear data, which is illustrated in Figure 3.9. The graph on the left-hand side shows that the feature raw frequency reaches a plateau at score level 4. Curvilinear regression analysis fits curves, rather than a straight line to predict the outcome by using polynomial equations. The term *quadratic trend* is used to describe one change in the direction of the line. A *cubic trend* is indicated when two changes in the direction (e.g., a rising-falling-rising trend) are observed. Table 3.7 illustrates the backward stepwise regression method used to arrive at the final model.

Table 3.7

Multiple Curvilinear Regression Modeling for Predicting Score Level

Variables	*B*	*Std error*	*Beta*	*t-value*	*p-value*	R^2	ΔR^2
Step 1						.303	.316
(Intercept)	3.06	0.04		75.09	< .001		
Raw feature frequency (linear)	11.22	1.44	0.41	7.76	< .001		
Raw feature frequency (quadratic)	-3.31	1.24	-0.12	-2.66	.008		
Raw feature frequency (cubic)	-1.13	1.14	-0.04	-0.99	.321		
Feature type (linear)	4.59	1.53	0.17	3.00	.003		
Feature type (quadratic)	-1.31	1.24	-0.05	-1.05	.293		
Feature type (cubic)	-0.48	1.02	-0.02	-0.47	.637		
Step 2						.323	.317
(Intercept)	3.06	0.04		75.14	< .001		
Raw feature frequency (linear)	11.18	1.44	0.41	7.75	< .001		
Raw feature frequency (quadratic)	-3.19	1.21	-0.12	-2.62	.009		
Raw feature frequency (cubic)	-1.29	1.08	-0.05	-1.19	.234		
Feature type (linear)	4.67	1.52	0.17	3.07	.002		
Feature type (quadratic)	-1.42	1.22	-0.05	-1.17	.245		
Step 3						.321	.316
(Intercept)	2.60	0.17		15.60	< .001		
Raw feature frequency (linear)	11.70	1.37	0.53	8.55	< .001		
Raw feature frequency (quadratic)	-3.89	1.05	-5.56	-3.70	< .001		
Raw feature frequency (cubic)	-0.71	0.96	-0.03	-0.74	.461		
Feature type	0.10	0.04	0.15	2.84	.005		
Step 4						.320	.317
(Intercept)	2.60	0.17		15.68	< .001		
Raw feature frequency (linear)	11.75	1.37	0.43	8.59	< .001		
Raw feature frequency (quadratic)	-3.91	1.05	-5.59	-3.72	< .001		
Feature type	0.10	0.04	0.15	2.80	.005		

The final model included a quadratic term for the feature raw frequency. $F_{(3)} = 216.25$, $p < .001$, $R^2 = .320$, $\Delta R^2 = .317$. Compared to the previous linear model, this curvilinear model had a statistically significantly improved fit ($F_{(1)} = 13.87$, $p < .001$).

In summary, the total frequency of the target features and the number of different feature types used in a text significantly predicted score level, explaining about 32% of score variance.

However, the relationship between the feature frequency and score level was non-linear.

3.3 The Use of Grammatical Features and Accuracy

To investigate whether there is an interaction between the patterns of grammatical features and accuracy in predicting grammar scores, I analyzed a subset of data (n = 196) that had been coded for errors. Table 3.8 shows the score level and topic distribution of this subset.

Table 3.8

Text Distribution of the Subset (n = 196) by Grammar Subscores and Topics

	Score level 1	**Score level 2**	**Score level 3**	**Score level 4**	**Score level 5**	**Total**
Topic A	8	16	18	13	10	65
Topic B	10	14	15	13	12	64
Topic C	12	14	16	15	10	67
Total	30	44	49	41	32	196

3.3.1 Descriptive statistics

For each text in subset of data, I examined the frequencies of target grammatical features, the number of feature types, and the EFCR. The mean values of these measures are presented in Table 3.9. The average frequency of any of the target features increased across score levels. The feature frequency showed the biggest increase between score levels 2 and 3, showing approximately 3 additional occurrences on average at score level 3. I found about 2.5 more occurrences at score level 4 compared to score level 3. Between score levels 4 and 5, the mean increase was smaller than 1. Regarding the number of different feature types used, texts at score level 1 included about three different types of features, with one additional type overserved at score level 2. The numbers slightly increased at the higher score levels, about 0.7, 0.3, and .05 more types added at each higher score level. The mean EFCR at each score level showed a steadily increasing trend.

Table 3.9

Description of Target Feature-Related Measures and Accuracy Measures

	Mean EFCR (SD)	Mean feature frequency (SD)	Mean number of feature type (SD)
Score level 1	.236 (.11)	5.43 (3.45)	2.87 (1.61)
Score level 2	.410 (.15)	7.11 (3.02)	4.00 (1.45)
Score level 3	.574 (.12)	10.40 (4.14)	4.71 (1.26)
Score level 4	.669 (.13)	12.90 (5.35)	5.00 (1.40)
Score level 5	.761 (.13)	13.50 (4.05)	5.34 (1.41)

Figure 3.11 visualizes the EFCR by score level (on the left-hand side). The 95% confidence intervals on the mean suggest that the mean EFCR for a score level is significantly different from that of each of the other levels.

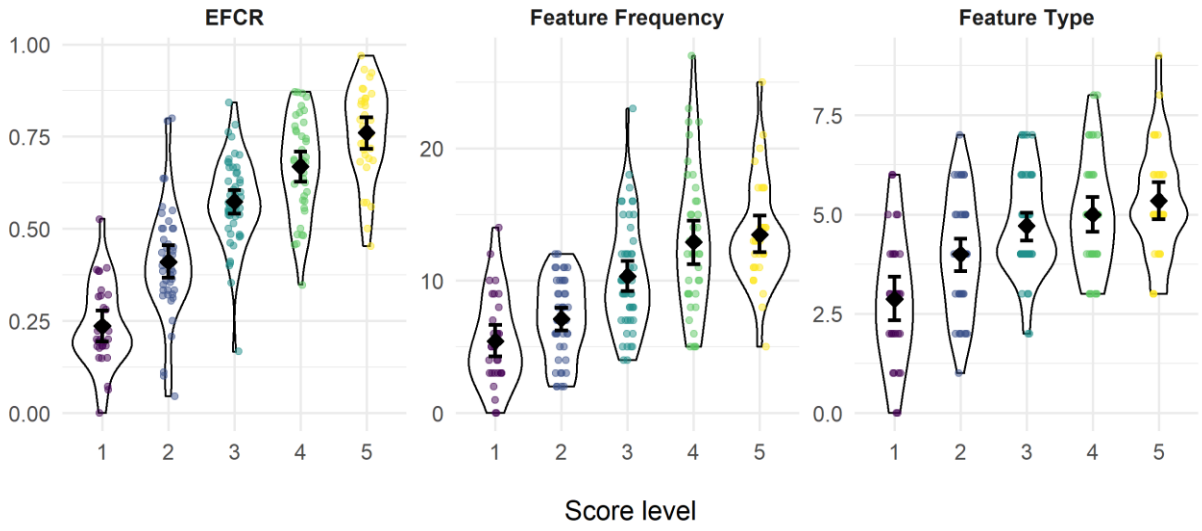

Figure 3.11. *EFCR, feature frequency, and feature type in the subset.*

3.3.2 Predicting score levels with regression analysis

To investigate (a) how well the accuracy measure (i.e., the EFCR) predict score level in addition to the use of grammatical features, and (b) whether there is an interaction between the two, I performed a multiple linear regression analysis.

Table 3.10

Regression Modeling for Prediction of Score Level

Variables	*B*	*Std error*	*Beta*	*t-value*	*p-value*	R^2	ΔR^2
Step 1						.668	.659
(Intercept)	0.423	0.333		1.27	.205		
Feature type	0.046	0.128		0.36	.719		
Feature frequency	0.012	0.054		0.22	.825		
EFCR	3.239	0.697		4.65	< .001		
Feature type * EFCR	0.094	0.209		0.45	.654		
Feature frequency * EFCR	0.049	0.082		0.59	.555		
Step 2						.668	.661
(Intercept)	0.340	0.276		1.23	.220		
Feature type	0.100	0.049		2.04	.042		
Feature frequency	-0.004	0.041		-0.11	.916		
EFCR	3.414	0.577		5.92	< .001		
Feature frequency * EFCR	0.074	0.059		1.25	.212		
Step 3						.665	.660
(Intercept)	0.079	0.181		0.44			
Feature type	0.078	0.047	.096	1.71	.089		
Feature frequency	0.043	0.016	.164	2.71	.007		
EFCR	4.031	0.300	.664	13.45	< .001		

The model, $F_{(3, 192)} = 214.54$, $p < .001$, $R^2 = .665$, $\Delta R^2 = .660$, explained about 66.5% of the variance in score level. Both the feature types employed and the EFCR were significant predictors of grammar score level. The model also showed that accuracy, as measured by the EFCR, had the largest effect on score level among the three predictors. No interaction effect was found between feature type and the EFCR or feature frequency and the EFCR. Although not presented here, I performed a regression that included a quadratic term following the previous finding that feature frequency showed a curvilinear trend, but found no resulting improvement in model fit.

3.4 Score Level Prediction

Although the analyses conducted thus far have provided insight into how well the use of grammatical features and accuracy predict score level on the CELC rating scale, they do not

inform how well these measures distinguish between adjacent score levels. Crucial decision making during both rater training and live rating often involves choosing between adjacent levels. Therefore, I performed a series of binary logistic regressions, predicting one category of a binary outcome, to examine how well the measures of accuracy and grammar use differentiate and classify adjacent score levels. I compared the four pairs of adjacent score level pairings: (a) score levels 1 and 2, (b) score levels 2 and 3, (c) score levels 3 and 4, and (d) score levels 4 and 5. In so doing, I performed four logistic regressions, each of which predicted the higher score of the two being compared. I report the logistic regression models as well as their performance in predicting and classifying texts into score levels in the following sections. For the predictor variables in each model, the odds ratio was computed. The odds ratio is "the change in odds of being in one of the categories of outcome when the value of a predictor increases by one unit" (Tabachnick & Fidell, 2013, p. 463). When the odds ratio is greater than 1, it means that a one-unit increase in the predictor is associated with increased odds of the outcome occurring. For instance, an odds ratio of 1.5 indicates that with a one-unit increase in the predictor, it is 1.5 times more likely that the outcome will be the target category. The model's effect size and goodness of fit are reported with Nagelkerke's R^2 and *C*-index. The *C*-index may be interpreted as the probability of a correct classification and ranges from .5 (indicating chance prediction) to 1.0 (indicating perfect prediction)

In these analyses, I used three predictor variables—the EFCR, feature frequency, and feature type—to assess how well these variables predict the grammar subscore. All variables were kept in each model even if a particular variable did not turn out to be a statistically significant predictor. In addition, removing a statistically non-significant variable did not improve the model fit in all cases. Each model's performance in predicting the score level and

classifying texts into score levels is reported after the corresponding regression analysis.

3.4.1 Logistic regression and validation: Score levels 1 and 2

The number of observations for this model was 74, with 30 texts rated at score level 1, and 44 texts rated at score level 2. The EFCR (accuracy) and number of feature types used were significant predictors of the score level: $\chi^2_{(2)} = 32.94$, $p < .001$, $R^2 = .485$, $C = .877$. As illustrated in Table 3.11, a text was 3.19 times more likely to be score level 2 with a 10% increase in the EFCR. With each additional type of grammatical feature used, a text was 2.24 times more like to be rated at score level 2.

Table 3.11

Logistic Regression Results for Score Level 2 Prediction

	Odds ratio	***95% CIs***	***Estimate***	***Std error***	***Wald Z***	***p-value***
Intercept			-4.323	1.204	-3.49	< .001
EFCR	3.19	1.89, 6.20	11.585	3.064	3.78	< .001
Type	2.24	1.08, 3.71	0.805	0.337	2.39	.017
Frequency	0.76	0.58, 1.05	-0.277	0.156	-1.77	.076

Figures 3.12 through 3.15 illustrate the probability of a text receiving the score being predicted (e.g., the probability of being score level 2 for Figure 3.12) for each of the three predictors. The top-left panel shows the probability (indicated on the y-axis) of a text being rated at the predicted score level with changes in the EFCR (indicated on the x-axis) when the other variables (i.e., feature types and frequencies) are held at average. The tick marks (called rugs) at the bottom of each graph represent individual observations in the dataset. The graph in the top-right panel of each figure illustrates the changes in the probability of a text receiving the higher score as the number of feature types changes. Likewise, the bottom-left panel shows predicted changes in probability of being scored at the higher of the two levels with changes in feature frequencies. These graphics are useful because they depict the varying degree of change in

probability depending on the value of the predictor.

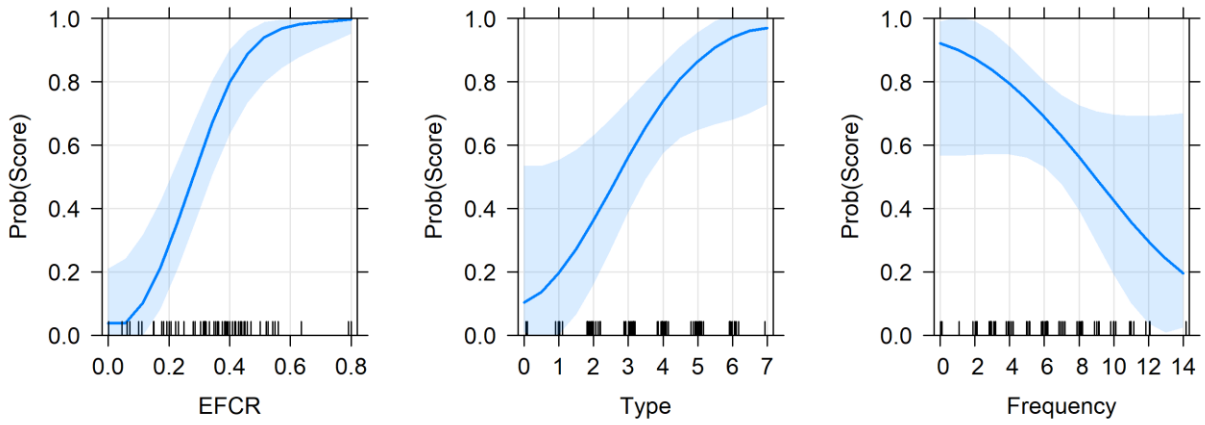

Figure 3.12. *Score level 2 probability as predicted by EFCR and feature types (mean EFCR = .34, mean type = 3.34, mean frequency = 6.43).*

For this model predicting score level 2, the EFCR and feature type were statistically significant. The effect plot for the EFCR (top-left panel) showed a steep increase in the probability of a text being rated at score level 2 as the EFCR increased. When a text had an EFCR of approximately .30, the probability of that text being rated at score level 2 was 50%, provided that the feature type and frequency had average values (3.34 types and 6.43 occurrences). When the EFCR increased to .40, the probability of a text being rated at score level 2 rose to 80%, indicating a large effect of the EFCR on score prediction. The effect size was smaller for feature type (plotted in the top-right panel), as signified by the relatively gentler slope. The probability of a text being rated at score level 2 became higher than 50% when three or more feature types were used. From there, using one additional feature type increased the probability to approximately 75%.

Table 3.12

Prediction and Classification Table for Score Level 2 Prediction

		Model prediction		Model validation	
		Score level 1	Score level 2	Accuracy	.851
Data	Score level 1 (n = 30)	23	7	Precision	.851
	Score level 2 (n = 44)	4	40	Recall	.909

Table 3.12 reports how well the logistic regression model predicts and classifies the texts in the data. The prediction based on the model is in columns (score level 1 and score level 2 under the model prediction column) and the actual data is in rows. The model was assessed on three measures: accuracy, precision, and recall. There were 30 texts in the dataset that were score level 1, and 44 texts that were score level 2 (as indicated in rows). Among these 30 texts at score level 1, the model predicted that 23 were score level 1 (the upper-left cell highlighted in gray) and 7 were score level 2. Among the 44 texts at score level 2, the model correctly predicted the score level of 40 texts (the lower-right cell highlighted in gray), but incorrectly predicted the scores of 4 of the texts to be level 1.

Model accuracy was computed by dividing the number of correct classifications (i.e., 23 + 40 in this example) by the total number of observations. Here, 63 out of 74 texts were correctly classified, representing 85.1% accuracy. Model precision answers the question: Out of all score level 2 predictions, what proportion was truly at score level 2? Out of 47 that were predicted to be score level 2 (the sum of score level 2 column in Table 3.12), 40 were actually rated as score level 2; therefore, the model showed 85.1% precision. Recall rate answers the question: Out of all data to be predicted (in this case, level 2 texts), what proportion was correctly predicted? Among the 44 level 2 texts, 40 were correctly predicted, representing a recall rate of 90.9%.

Taking all the above measures into account, the logistic regression model for predicting score level 2 was effective with the EFCR and number of feature types as its predictor variables.

3.4.2 Logistic regression and validation: Score levels 2 and 3

I found the EFCR and feature frequency to be significant predictors in the logistic regression model predicting score level 3, $\chi^2_{(3)} = 36.53$, $p < .001$, $R^2 = .434$, $C = .856$. With a 10% increase in EFCR, a text was 2.3 times more likely to be rated at score level 3. When feature frequency increased by 1 in a given text, the text was about 1.3 times more likely to be rated at score level 3.

Table 3.13

Logistic Regression Results for Score Level 3 Prediction

	Odds ratio	***95% CIs***	***Estimate***	***Std error***	***Wald Z***	***p-value***
Intercept			-5.473	1.412	-3.86	< .001
EFCR	2.31	1.53, 3.78	8.355	2.291	3.65	< .001
Type	0.89	0.50, 1.53	-0.122	0.280	-.044	.662
Frequency	1.26	1.01, 1.61	0.229	0.116	1.97	.049

For texts at score levels 2 and 3, the probability of being score level 3 crossed the 50% line when the EFCR was around .45 and feature frequency was 8. The probability steadily increased with higher EFCR and higher feature frequency.

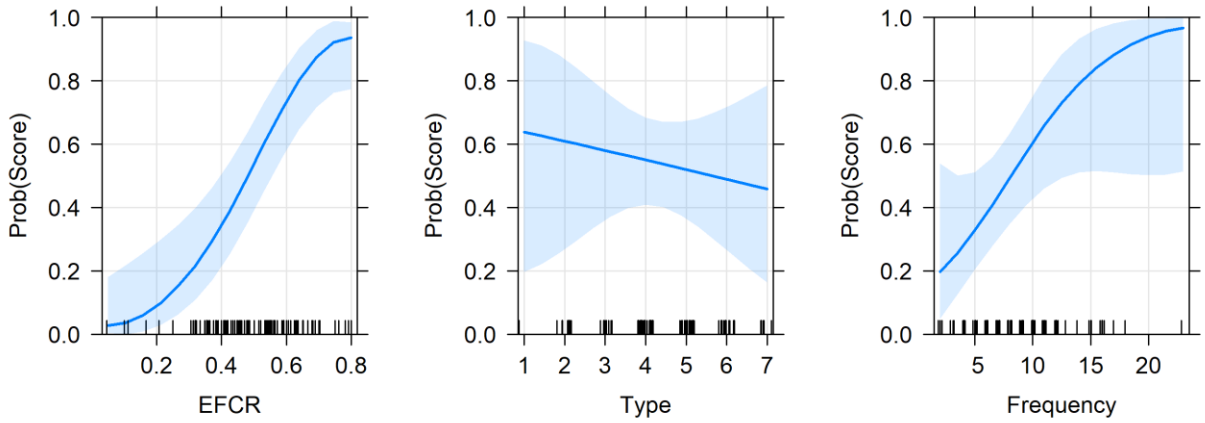

Figure 3.13. Score level 3 probability as predicted by EFCR and feature types (mean EFCR = .50, mean type = 4.38, mean frequency = 8.82).

When a text had an EFCR of approximately .50, the probability of that text being score level 3 crossed the 50% chance line. Additional increases in the EFCR were accompanied by commensurate increases in the probability of a text being rated at score level 3. For feature frequency, the data points mostly ranged from 4 to 12. The predictive power of feature frequency outside of this range is unstable, as indicated by the large shaded area around the line (95% confidence intervals). The probability of a text being score level 3 rose above 50% when the target grammatical features appeared in a text 8 times.

Table 3.14

Prediction and Classification Table for Score Level 3 Prediction

		Model prediction		Model validation	
		Score level 2	Score level 3	Accuracy	.785
Data	Score level 2 (n = 44)	35	9	Precision	.809
	Score level 3 (n = 49)	11	38	Recall	.776

The model predicting score level correctly predicted approximately 78.5% of the data. Score level 3 prediction were accurate 80.9% of the time and 77.6% of the text that received Score 3 were correctly classified as score level 3.

3.4.3 Logistic regression and validation: Score levels 3 and 4

For the 90 texts at score level 3 (n = 49) and 4 (n = 41), only the EFCR served as a significant predictor in the logistic regression model, $\chi^2_{(3)} = 17.43$, $p < .001$, $R^2 = .235$, $C = .750$. The goodness-of-fit statistics indicated that this model was not as effective as the previous ones.

Table 3.15

Logistic Regression Results for Score Level 4 Prediction

	Odds ratio	***95% CIs***	***Estimate***	***Std error***	***Wald Z***	***p-value***
Intercept			-5.433	1.702	-3.19	.001
EFCR	1.85	1.27, 2.85	6.162	2.047	3.01	.003
Type	1.06	0.68, 1.63	0.054	0.221	0.25	.806
Frequency	1.10	0.98, 1.26	0.100	0.062	1.60	.109

With a 10% increase in accuracy, a text was 1.85 times more likely to be score level 4. This effect size was smaller than that overserved in the score level 3 prediction. The other two features were not statistically significant.

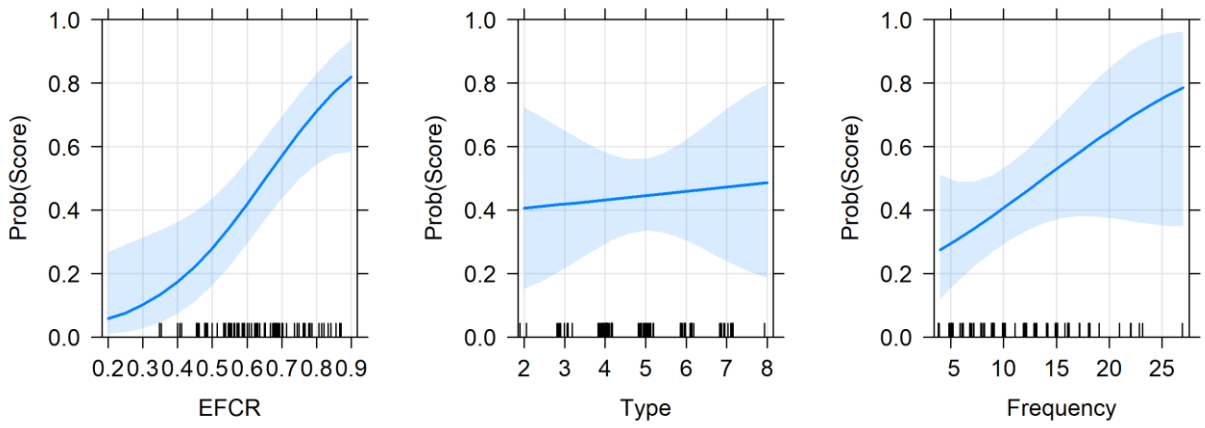

Figure 3.14. *Score Level 4 probability as predicted by EFCR and feature types (mean EFCR = .62, mean type = 4.84, mean frequency = 11.52).*

The plot for the EFCR shows a much gentler slope and the highest predicted probability is lower compared to the two previous models. When a text showed an EFCR over .65, it was more likely for the text to have been rated at score level 4.

Table 3.16

Prediction and Classification Table for Score Level 4 Prediction

		Model prediction		Model validation	
		Score level 3	Score level 4	Accuracy	.689
Data	Score level 3 (n = 49)	37	12	Precision	.676
	Score level 4 (n = 41)	16	25	Recall	.610

The accuracy of the score prediction was 68.9%. Of the 37 texts predicted to be Score 4, 67.6% were indeed rated as Score 4 by human raters. Of the 41 texts rated at score level 4, 61.0% were correctly classified with this model.

3.4.4 Logistic regression and validation: Score levels 4 and 5

Regression modeling to predict score level 5 versus score level 4 was performed on 73 texts. When the EFCR and the number of types of features were included in the logistic regression model, only the EFCR was found to be a significant predictor of actual score level: $\chi^2_{(3)} = 10.55$, $p = .015$, $R^2 = .180$, $C = .708$.

Table 3.17

Logistic Regression Results for Score Level 5 Prediction

	Odds ratio	***95% CIs***	***Estimate***	***Std error***	***Wald Z***	***p-value***
Intercept			-5.824	2.006	-2.90	.004
EFCR	1.80	1.23, 2.82	5.900	2.099	2.81	.005
Type	1.30	0.87, 1.99	0.262	0.207	1.26	.207
Frequency	1.00	0.89, 1.12	-0.003	0.058	-0.01	.995

The EFCR was still a significant predictor in score level prediction. When other variables were held at average, a 10% increase in EFCR would make a text 1.8 times more likely to have been rated at score level 5. This effect size was similar to Score 4 prediction, but the model fit was inferior.

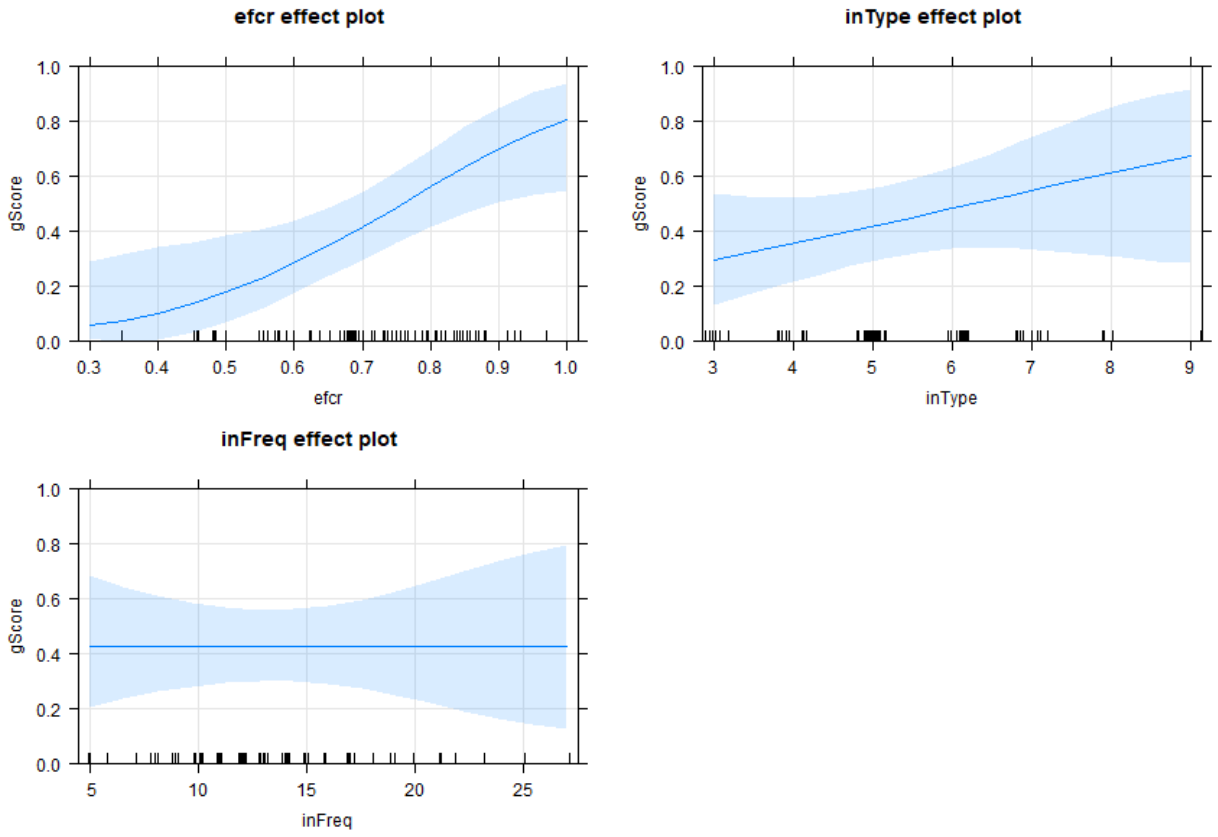

Figure 3.15. *Score level 5 probability as predicted by EFCR and feature types (mean EFCR = .71, mean type = 5.12, mean frequency = 13.10).*

Table 3.18

Prediction and Classification Table for Score Level 5 Prediction

		Model prediction		Model validation	
		Score level 4	Score level 5	Accuracy	.630
Data	Score level 4 (n = 41)	30	11	Precision	.593
	Score level 5 (n = 32)	16	16	Recall	.500

The model accurately predicted the scores for 63.0% of the texts. Among score level 5 predictions, 59.3% were correct. Of the 32 texts rated to be at score level 5, only 50% were correctly classified as level 5 by the model.

To summarize, logistic regression models with the measures of feature use (as measured by feature frequency and feature type) and linguistic accuracy (as measured by the EFCR) as

predictor variables distinguished between two adjacent levels with varying degrees of confidence. Predictions at lower score levels, i.e., score levels 2 and 3, had a higher probability of being accurate. The EFCR emerged as a significant predictor for all level predictions. Feature type and feature frequency significantly contributed to prediction of score level 2 against 1 and of score level 3 against 2, respectively. In other words, adopting a larger number of grammatical features distinguished score level 2 from score level 1, and using the features more frequently characterized score level 3. Overall, the characteristics regarding grammatical feature use showed more variance among score levels 1 through 3, and thus were more useful in distinguishing between the lower levels. A text's level of accuracy also showed more predictive power at these lower levels than at the higher levels.

CHAPTER 4: DISCUSSION

The aim of this study was two-fold: to investigate the use of specific grammatical features that would help reveal characteristics of different score levels in L2 English writing; and to find validity evidence for test score interpretation by examining how well the rating rubric reflects actual written responses. In this section, I discuss my findings in relation to the research questions, previous research, and research contexts that outlined in previous chapters.

4.1 Patterns of Target Grammatical Features Use and Score Levels

4.1.1 Use of target grammatical features

In the first part of the analysis, I investigated the overall use of the target grammatical features by surveying the frequencies of features attested in the CELC corpus. The most frequently used features were post-nominal modification with relative clauses, *that*-complement clauses controlled by verb (e.g., I *believe that* teachers can make learning more fun), and *to*-

clauses controlled by verb (e.g., I *want to go* to Italy). These three features also appeared in a relatively larger proportion of the corpus, showing a wide distribution. Across score levels, the relative (normalized) frequencies showed statistically significant differences for several features: (a) relative clauses (between score levels 1 and 2 and score levels 2 and 3), (b) *to*-complement clauses controlled by verb (between score levels 1 and 2), (c) *to*-complement clauses controlled by noun (between score levels 2 and 3), and (d) *wh*-word pseudocleft constructions (e.g., I don't know *what he wants*; between score levels 2 and 3).

One of the merits of using specific grammatical features in this study was the availability of *English Grammar Profile*, a resource that specifies which features are characteristic of which CEFR level. As introduced in Table 2.5, relative clauses and *wh*-word pseudoclefts are identified as a B1-level feature, *to*-clauses controlled by verb a are B1- or B2-level feature (depending on the type of the verb), and *to*-clauses controlled by noun are a B2-level feature. As a reminder, the B2 level ability is determined at score level 3 above for the CELC. The empirical findings in this study demonstrate that the use of these four features representative of the B1 and B2 levels show significant differences in scoring levels 1 through 3. The proportions of the texts including these features were also relatively large (ranging from 30.7% to 85.7%). Taken together, these results suggest that these four features are visible and useful characteristics which can aid the determination of B2-level grammar ability.

These results shared some similarities and dissimilarities with Biber et al.'s (2016) findings, where the researchers reported how well the mode (i.e., speaking or writing), task type (i.e., TOEFL iBT independent and integrated), and score level (ranging from 1 to 4) predicted frequencies of 23 grammatical features. Whereas Biber et al. (2016) found score level to be significantly associated with verb + *that*-complements, I did not find any statistically significant

differences in the frequencies of this feature across score levels. In their study, Biber et al. reported that task type and score level together predicted the frequencies of desire verb + *to*-clauses (e.g., *want* to go) and noun + *to*-clauses. In the current study, the frequencies of verb + *to*-clauses at score level 1 were found to be significantly lower than at the higher levels, but there were no significant differences among higher levels (e.g., no difference between score levels 2 and 5). The frequencies of noun + *to*-clauses significantly distinguished score levels 1 and 2 from score levels 3 and above (refer to Appendix D). However, because Biber et al. did not report the frequencies of the features investigated or at which levels each feature showed differences, it is difficult to grasp how exactly the patterns compare.

Staples et al. (2016) also examined the use of specific grammatical features, many of which overlapped those that were used in this study. In their study, L1 English academic writing was evaluated by comparing grammatical complexity development across levels of study (i.e., first-year undergraduate, second-year undergraduate, final-year undergraduate, and graduate). The features that significantly predicted levels of study included *wh*-word complement clauses, verb + *that*-clauses, noun + *that*-clauses, and relative clauses. The frequencies of all four clause-level features markedly decreased between Level 3 (final-year undergraduate) and Level 4 (graduate). What this study illuminated was that complex, clausal features are used more frequently at lower levels (among L1 English speakers) than at the higher level.

One notable finding in the study of syntactic development of L2 writing that relates to rating was revealed in Crossley and McNamara's (2014) longitudinal study. The researchers found that, developmentally (i.e., comparing the essays written at the beginning of a semester and at the end of the semester), phrasal complexity, especially of noun phrases (e.g., number of modifiers per noun phrase, number of words before the main verb) showed significant changes.

However, when predicting both language use score (assigned as a category subscore) and overall essay score, cross-sectionally, measures related to clausal complexity (i.e., incidence of all clauses, *to*-infinitives, and *that*-complements with verbs) were found to be significant predictors. In other words, control of clausal features was a better predictor of rater judgment than phrasal complexity. This may be due to fact that the rubric descriptors (an analytic rubric with categories of content, organization, vocabulary, language use, and mechanics) did not specify phrasal-level language control. As with the CELC rubric descriptors for the grammar category, the rubric used in Crossley and McNamara's (2014) study referred explicitly to syntactic variety in its language use category descriptors, but not necessarily to phrase-level ability. The finding, therefore, suggests that phrasal complexity may need to be considered as part of a rating rubric to more accurately capture development in L2 English writing ability.

To summarize, the findings in existing literature suggest that complex clausal-level features develop over lower to intermediate levels of English proficiency. In the current study, I also located significant differences in the frequencies of certain grammatical features only among the lower levels. Overall, the use of 14 grammatical features was moderately useful in evaluating B2-level (intermediate) grammatical ability: The frequencies of the 14 features and the number of different types used in the texts accounted for about 32% of the variance in the grammar subscore. However, to better account for the difference among higher score levels, more measures may be needed, and I return to this argument when I consider the implication of this study.

4.1.2 Co-occurring grammatical features

Previous research indicated that essay quality can be better explained by a group of linguistic features that occur together than by the use of individual features (Biber et al., 2016;

Jarvis et al., 2003; Yan & Staples, 2016). To identify the co-occurring patterns of target grammatical features in this study, I performed a principal components analysis which clusters highly correlating variables that account for the variance in the data. Of the six components identified, five were significantly associated with grammar subscore. To review, the use of relative clauses and nouns + *to*-complements contributed the most to the first component. These two features share the same functional characteristic, in that they create complex noun phrases with post-nominal modifications. This suggests that the ability to formulate complex nouns indicates higher grammatical ability. The second component was characterized by the use of verb + *that*-complements, *wh*-word finite pseudoclefts, and less use of prepositional dative constructions. This component can be interpreted as the ability to use finite dependent clausal features. The positive correlation between less use of the prepositional dative and score levels may suggest that the use of non-clausal features negatively associated with grammar ability. However, I observed that many instances of the prepositional dative construction included errors, namely, incorrect preposition (e.g., "I told my secret *in* my friend"), and this may have contributed to the negative association with the score. The third component included many of the *to*-complement features, reflecting the use of non-finite dependent clauses. Component 4 indicated the ability to use complex *to*-complements, *to*-complement subject-to-object raising constructions with verbs (e.g., "I *found the task to be* difficult") and with passive verbs (e.g., "He *was known to tell* lies"). Finally, the use of ditransitive constructions was a significant predictor, albeit having the smallest effect size of all five significant components.

Table 4.1

Characteristics of Co-occurring Features

Component	Contributing features	Biber et al.'s (2011) categorization	Significance test
Component 1	• relative clauses • nouns + *to*-complements	finite noun modifiers	Score 1 < Score 2 < Score 3 < Score 4
Component 2	• verb + *that*-complement • *wh*-word finite pseudocleft • less use of prepositional dative construction	finite complement	Score 1 < Score 2 < Score 3
Component 3	• *it* extraposed adjective + *to*-complement • *wh*-word + *to*-complement	non-finite complement	
Component 4	• verb + object + *to*-complement (subject-object-raising) • verb passive + object + *to*-complement	non-finite complement	Score 2 < Score 3
Component 6	• ditransitive clause	-	Score 1 < Score 2

The co-occurring feature patterns observed in this study largely coincided with the categorization of grammatical features laid out by Biber et al. (2011). The authors distinguished three grammatical types: finite dependent clauses, non-finite dependent clauses, and dependent phrases. Each type was subdivided depending on the function of the feature: adverbial, complement, and noun modifiers. Component 1 falls into the finite noun modifier subcateogry, Component 2 into the finite complement subcategory, and Components 3 and 4 into the non-finite complement subcateogry. As such, the analysis and results in this present study revealed that the actual patterns of grammatical feature use can be mapped onto the "complexity devices" that Biber et al. (2011) identified based on previous corpus-based studies.

Existing studies that investigated co-occurring patterns of specific grammatical features have focused on revealing differences in language use by mode, task type, genre, L1, and/or

proficiency. In this line of multidimensional analysis research, factors such as which linguistic features were included as variables and how many variables were analyzed significantly impact the identification of dimensions (more commonly known as factors in factor analysis). Therefore, unless an identical same set of measures are used, each of these analysis stands on its own. What was interesting in the results of this study was that the investigated features patterned into certain structural categories. Although further investigation is needed, this may serve as empirical evidence that grammar structures of similar types (e.g., finite, non-finite) and functions (e.g., complement, noun modifier) occur together and exhibit (cross-sectional) developmental patterns.

4.1.3 Use of the target grammatical features and relationship to score levels

In this study, the characteristics of target feature use were operationalized by the number of different types of features used and the frequencies of the features. These two characteristics were useful in understanding the relationship between the score levels and the patterns of target feature use attested in the learner-produced texts. Interestingly, regression analysis did not show relative frequency of the target features to be a significant predictor of score level. In the field of corpus linguistics, it is advised to normalize frequencies when comparing texts of varying lengths (Biber, Conrad, & Reppen, 1998). However, as the score level was assigned to each text by human raters, interpreting the raw frequency may make more sense in understanding the characteristics of feature use in relation to score level. The fact that raw feature frequency was significant factor also suggests a possible interaction between grammar use and fluency, as the raw frequency of feature occurrence is contingent on how long the text is. Previous research into textual features has consistently found that text length significantly correlates with text quality (e.g., Banerjee et al., 2015; Yan & Staples, 2016). Because the present study investigated how the use of specific grammatical features attested in an exam response associated with score, I

found the measures that more holistically reflect the text to be more effective (i.e., raw frequency per text as opposed to relative frequency). Similarly, the total number of different types of features used in each text had a stronger effect than the six component scores, each of which characterized a specific aspect of feature use. Although identifying the components explained how certain grammatical features tend to co-occur, they did not turn out to be significant predictors of score levels when entered into regression models along with the measures of feature frequency and feature type. The two more holistic measures, therefore, seem more helpful in predicting score levels. The finding that the number of different types of features is a significant predictor of score levels corroborates Park's (2017) study in which she found the number of verb-argument types used in an essay was most predictive of essay score, over traditional measures of syntactic complexity (e.g., mean length of T-unit, dependent clauses per T-unit, complex noun phrases per T-unit).

In sum, two variables—number of types (range) and frequency—explained about 32% of the variance in the grammar subscore. This moderate effect size suggests that syntactic complexity and range are being evaluated by the existing CELC rating rubric. However, it should be noted that the relationship between feature use and score level was not entirely linear, with non-linearity shown at the two highest score levels. This may be due to the fact that the highest score level included texts that received an average of 4.5 and were thus not truly representative of a score of 5. However, considering that score level 1 also consisted of texts with an average score of 1.5 (therefore close to score level 2), it could be argued that the difference in syntactic range is more prominent at the lower proficiency levels. This phenomenon has been pointed out by O'Keeffe and Mark (2017), who argued that structural complexity progresses at the lower proficiency levels, and it is the lexical range with which grammar structures are used that

develops at higher proficiency levels.

4.2 Error and Syntactic Variety

The EFCR significantly differed across score levels. This finding is consistent with previous research that examined any type of measures of accuracy or errors (e.g., Alexopoulou et al., 2017; Banerjee et al., 2007; Cumming et al., 2005; Thewissen, 2013; Verspoor et al., 2012). Accuracy measure was also a significant predictor in distinguishing each pair of adjacent levels; however, its effect was greater for predicting the lower score levels (i.e., odds ratios of 3.19 and 2.85 for predicting score levels 2 and 3, respectively) and smaller at higher score levels (i.e., odd ratios of 1.85 and 1.80 for predicting score levels 4 and 5). This is somewhat different from Knoch et al.'s (2014) study where the researchers found no significant differences in accuracy (as measured by EFCR) between the two lowest levels (on a scale of 1 to 5). However, the lowest group's EFCR was around .40 in their study whereas the lowest score level in this study averaged to approximately .25. For the higher score levels, Knoch et al. (2014) found significant differences, except for on the independent task in one of the two parallel test forms. Their findings, however, need cautious interpretation as the sample sizes for the highest (n = 6) and the second highest levels (n = 59) were substantially different. The current study included more comparable sample sizes for the score levels, but again, the highest score level was an aggregate of grammar subscores 4.5 and 5, rather than a clear 5.

Verspoor et al. (2012) reported a significant decrease in the number of lexical and spelling errors at the lower score levels (defined as CEFR A1-A2 levels) and a significant decrease in the number of total errors across levels A1 to B1. In their study, grammar errors (e.g., singular/plural, L1 word order, incorrect word form, L1 constructions) attested too infrequently for a meaningful conclusion to be drawn. Thewissen (2013) examined a range of CEFR

proficiency levels, from B1 to C2 and found that, among the features that showed development in accuracy as proficiency level progressed, most of the improvement occurred between the B1 and B2 levels. The current study similarly finds that the accuracy measure was a stronger predictor of score level at the lower bands. Interestingly, in both Verspoor et al. (2012) and Thewissen's (2013) studies, the grammar or clausal-level error types occurred infrequently and did not show developmental patterns. Similarly, in Alexopoulou et al.'s (2017) study on the relationship between proficiency level and accuracy at the morphological or lexical level, the results clearly showed a decrease in number of errors as proficiency (ranging from CEFR A1 level to the C2 level) increased.

In the regression analyses I performed, I found no interaction between accuracy and feature range or frequency. The lack of apparent interaction between clausal-level features and errors on score levels may be due to the fact that the error coding was more concerned with morphological errors, which all aforementioned studies found to show significant improvement as proficiency progresses, especially at the lower proficiency levels. In the current study, most of the error types coded also could be attributed to morphological or lexical errors, and I found that accuracy (i.e., the EFCR) consistently increased across score levels. Although I utilized the measure of EFCR to operationalize grammatical accuracy described in the rating rubric, the EFCR does not distinguish between different error types. Therefore, the quality of each type of feature used requires further investigation to fully comprehend how the interaction between the use of a specific grammatical feature and its accuracy affects score level. This point of limitation will be further discussed in Section 5.2.

4.3 Evidence for CELC Test Validity

I examined two assumptions that would lend support to the validity of the CELC rating

rubric descriptors and scale: (a) examinees' texts reflect the descriptions of performance in the rating rubric, and (b) linguistic features of examinees' texts differentiate between score levels. The first assumption was explored by identifying complex clausal features, as appear in rubric descriptors: "multi-clausal sentences and syntactic variety to effectively clarify, explain and elaborate" (score level 5); "control of a range of syntactic forms that allows writer to efficiently and effectively convey meaning and ideas" (score level 4). The frequency of error indicated in the rubric descriptors was operationalized by annotating errors and computing the EFCR. When feature frequency, feature type, and EFCR were used to predict the grammar subscore, I found feature frequency and EFCR to be statistically significant predictors. The model explained about 66% of the variance in the grammar subscore, with the EFCR substantially explaining the score. I further examined each pair of adjacent levels to find support for the assumption that these features differentiate between score levels. In all cases, the EFCR was a significant predictor.

Between score levels 1 and 2, the number of different types of features used contributed to distinguishing these two levels. Between score levels 2 and 3, the frequency of features significantly predicted score level 3 along with the EFCR. For the models predicting scores at level 4, only the EFCR was a significant predictor. More crucially, differentiating score level 4 from score level 3 with the features investigated in this study and the EFCR was not very successful, indicating a potential weakness in the rubric and rater training materials and protocols. The rubric descriptors express that while score level 3 shows control of "basic syntactic forms," score level 4 shows control of "a range of syntactic forms." In other words, texts at score level 4 are expected to exhibit progress in syntactic range over those at score level 3 according to the rubric, but the investigation into the 14 grammatical features showed no significant differences. This finding calls for further investigation into syntactic variety at these

two levels. Based on the previous discussion, one possible explanation is that above the B2-level is when lexical diversity, pragmatic understanding, and knowledge of collocations progress more than grammatical repertoire (O'Keeffe & Mark, 2017). Another important point raised by a series of studies (Biber et al., 2011; Biber et al., 2016; Staples et al., 2016) is that phrasal complexity (e.g., number of modifiers per noun phrases, use of attributive adjectives, noun phrasal length, verb phrases, and prepositional phrases) shows more development at the higher levels of proficiency.

Because the rubric descriptors depict the relationship between syntactic range and control (i.e., "effectively," "efficiently," and "adequate[ly]" using various syntactic forms), I expected to find some interaction between these two elements. For all level predictions, I found no interaction between feature use and accuracy. However, I acknowledge that the EFCR as a measure for accuracy pertains to the overall accuracy of the texts and not necessarily the effectiveness of the grammatical features of the texts. Therefore, inspecting the effectiveness of grammar use across score levels would require examining the accuracy and effectiveness of each feature occurrence. For further discussion on the implication of the findings for the CELC, I turn to the next chapter.

CHAPTER 5: CONCLUSION

5.1 Implications for Testing and Research

This study has direct implications for CELC rating as well as for test validity research in general. I examined the relationship between rating rubric descriptors for grammar use and the grammatical features attested in examinee-produced texts. To review, the CELC rubric descriptors for the grammar category largely identify syntactic range and control (e.g., *control of*

basic syntactic forms, control of a range of syntactic forms, effective use of multiclausal sentences and syntactic variety), describing a progression of syntactic forms from basic to advanced, and of accuracy from more errors to fewer errors. These constructs were operationalized as the frequency of occurrence of 14 grammatical features, the number of different types of features used, and the EFCR. The 14 selected grammatical features were relevant to the context in two respects: They were described in empirical studies on CEFR level descriptors (English Profile, 2015; Hawkins & Filipović, 2012) and in empirical studies on writing development (e.g., Biber et al., 2011; 2013). However, these features were not an exhaustive list of all possible grammatical structures and therefore may not reflect the full range of structures that constitute the construct of syntactic range. With that in mind, the results indicated that the current rubric descriptors matched the texts according to the rater-assigned score level at the lower levels (score levels 1 through 3) but not so much at the higher levels (score levels 3 through 5). The reason for this result can be attributed to a number of issues discussed in the previous section. Nevertheless, because the CELC certifies proficiency at the CEFR B2 level (i.e., score level 3 and above), it can be argued that the rubric descriptors are reflective of actual examinee-produced texts to a certain extent; that is, the constructs described in the rubric (i.e., syntactic range and control) manifest differently at the lower score levels including between score levels 2 and 3.

From an L2 writing development perspective, this may be an expected outcome as previous research suggests that the use of grammar structure may not show substantial development at high proficiency levels. However, for the purpose of assessment, more linguistic features may need to be included in the rating rubric and rater training materials. For instance, Biber and his colleagues (e.g., Biber et al., 2011; Biber & Gray, 2013) have argued that phrasal

complexity development is characteristic of academic writing and language use at higher proficiency levels. In the future, phrasal complexity, such as the use of complex noun phrases with longer modifiers and number of words before the main verb, can be investigated to empirically examine whether these features help distinguish the higher score levels. If found to be the case, depiction of such characteristics could be included in the CELC rubric descriptors, as well as in rater training materials, to provide stakeholders with fuller information about score interpretation and more reliable rating processes.

For the broader field of language testing, this study offers an example of monitoring rating materials to better support the validity of score inferences. As Banerjee et al. (2015) stated, rating scale development and revision can benefit from combining "our current understanding of the indicators of second language writing development" and "the empirical analysis of performance data (p. 6)." Test developers and researchers are expected to constantly maintain and revise testing materials, so that test users can make clear inferences to constructs the test intends to measure. In this study, I approached test validity by examining the relationship among rating rubric descriptors, scale, and examinee-produced texts. My findings were somewhat inconclusive in that investigating syntactic range with 14 grammatical features might not have been comprehensive enough to fully operationalize the construct of syntactic range and its effectiveness. Although the use of these specific features significantly differed at the lower score levels, including more types of grammatical features would provide a fuller picture of what examinees can do with grammar and whether there exist significant differences at higher score levels. As discussed earlier, expanding the grammatical range to incorporate phrasal complexity may also be useful in supplementing the current rubric descriptors to add more cues for distinguishing between the grammatical range of higher-level texts.

In summary, the findings from this study highlight the need for empirically investigating how well the constructs of rubric descriptors attest in test performance data. What has been conceptually described by experts, such as the CEFR level descriptors (Council of Europe, 2018), requires validation in their specific context to ensure valid score inference. In addition, writing assessment materials may benefit from consideration of the tangible characteristics of L2 development found in writing development research, such as the range of phrasal complexity (Biber & Gray, 2013; Crossley & McNamara, 2014), morphological accuracy (Alexopoulou et al., 2017; Thewissen, 2013), and association strength (i.e., the degree of idiomaticity) between a construction and its lexis (e.g., Kyle & Crossley, 2017; O'Keefe & Mark, 2017), which demonstrate different profiles across different proficiency levels.

5.2 Limitations and Future Directions

In this study, I touched upon the issues of test validity with regard to rating processes for L2 English writing and L2 English grammar development. There are a number of limitations to be addressed in this study in order to advance this line of research. One point for caution in interpreting the results is that the score levels described here only pertain to grammar and not overall writing quality or proficiency. The benefit of this design was that I was able to isolate the raters' judgments about grammatical ability from overall writing quality or proficiency, thereby excluding other constructs. However, most studies in writing assessment and writing development have considered overall writing quality. Although the grammar subscore and writing score showed significant, high correlation for the CELC (N = 3,334, r = .914, 95% CIs [.909, .920], $p < .001$), the findings of this study technically pertain to grammar category scores, and hence grammatical ability.

Another point to be noted is that because the CELC is administered in Greece, the

examinees mostly consist of L1 Greek learners of English. Although this does not pose any issues for the purpose of validating the local rating process, I do not claim the linguistic characteristics revealed in this study to be universal. The results can serve as a good addition to the Cambridge Learner Corpus data or as a point of comparison to previous studies profiling features of L2 English.

As noted, one limitation in investigating the relationship between rating rubric descriptors and score levels was that the "adequacy," "effectiveness," and "efficiency" indicated in the descriptors were not well operationalized in this study. The error coding guidelines used in this study included error types that are better characterized as grammatical errors than morphological errors, such as missing subjects, missing verb complements, problems with relative clauses, and unclear structures. One direction for future research is to group error types into categories such as lexical, morphological, and grammatical; however, determining the error type in this way is not always straightforward and is subject to the coders' thinking processes (Polio & Shea, 2014). To my knowledge, there is no study that has investigated the accuracy or acceptability of a list of grammatical features across proficiency levels or levels of writing quality. Future research can assess the effectiveness of grammar use in relation to score or proficiency level by evaluating the accuracy of various types of grammatical features attested in learner language.

Another limitation is that because the CELC is designed to be certify CEFR B2-level proficiency, variety in proficiency levels of test candidates may be restricted to a band surrounding the B2 level. More advanced L2 English learners are more likely to opt for a test that certifies higher-level proficiency, for instance, the *Certificate of English Language Proficiency*, which certifies CEFR C2-level proficiency. Although CELC score levels serve as an indication of grammatical ability and proficiency level, it should be noted that they do not

exactly correspond to specific CEFR proficiency level.

Lastly, the scope of this study was limited to syntactic features without consideration of the lexis used with these features. The testing context in this study utilized an analytic rubric which separately categorized grammar and vocabulary. Examining the grammar category was a valuable starting point for understanding and assessing the linguistic features of examinee texts; however, it is widely accepted that grammar use needs to be considered in combination with lexis (see Römer, 2009). The *English Grammar Profile*, for example, differently identifies the CEFR level of a grammatical construction depending on the lexical profile of the words used in the construction (e.g., *want* + *to*-complement as a B1-level feature vs. *fail* + *to*-complement as a B2-level feature). In addition, O'Keeffe and Mark (2017) reported increased complexity in terms of lexis with the same grammatical constructions at higher proficiency levels. Examining the level of lexical sophistication within grammatical features may prove useful for understanding and differentiating characteristics of different proficiency levels.

APPENDICES

APPENDIX A

MSU-CELC Essay Evaluation Rubric

(Michigan State University English Language Examinations, n.d.)

CELC Rating Scale	Linguistic Competency - Grammatical Accuracy	Linguistic Competency - Range of Vocabulary	Development and Task Completion	Genre Appropriateness and Writing Conventions
5 Characteristics of "Honors Pass" (exhibits some C1/C2 features)	Meets all B2 requirements for this category, plus effectively uses advanced structures such as multi-clausal sentences and syntactic variety to effectively clarify, explain and elaborate support for the point of view assumed in the essay	Meets all B2 requirements for this category, plus skillfully selects some less common words and uses a wide variety of lexical expressions to clarify and enhance the point of view assumed in the essay	Meets all B2 requirements for this category, plus achieves an exceptionally well-balanced and complete expression of point of view through elaboration, explanation and clarification of key points in a manner that is easy to follow	Meets all B2 requirements for this category, plus language and essay structure match the B2 task exceptionally well. Creative & skillful use of conventions such as sentence fragments for emphasis; mechanically excellent
4 Characteristics of "Clear Pass" (exhibits B2/C1 features)	Control of a range of syntactic forms that allows writer to efficiently and effectively convey meaning and ideas relevant to the B2 task; few errors	Control of wide range of vocabulary is precise enough to efficiently and effectively convey meaning relevant to the B2 task	Response to prompt is clear and skillfully and comprehensively supported by relevant description, examples, explanations and/or arguments	Full control of major conventions, strong control of genre and register
3 Characteristics of "Marginal Pass" (B2 "floor")	Control of basic syntactic forms is adequate to convey meaning and ideas relevant to the B2 task without causing confusion, even though some errors may be present	Control of vocabulary is adequate to convey meaning relevant to the B2 task without causing confusion, even though some inaccurate word choices occur	Response to prompt is clear in terms of viewpoint, which is adequately, though minimally, supported by relevant description, examples, explanations and/or arguments	Displays acceptable range of register and genre; shows appropriate sense of audience in use of conventions, which tend toward a standard format; some non-disruptive errors present
2 Characteristics of "Narrow Fail" (satisfies some, but not all B2 criteria)	Control of basic syntactic forms is NOT adequate to convey meaning and ideas relevant to the B2 task without causing confusion; numerous errors are present and limit effectiveness of the text	Control of vocabulary is NOT adequate to convey meaning relevant to the B2 task without causing confusion – weaknesses include incorrect word forms & word choices, limited range, repetition of words, repetition of prompt	Some ability to address prompt – Inadequate, limited or trivial analysis, some relevant arguments, may be repetitious, simplistic or exceedingly immature	Only some conventions followed or conventions followed inconsistently – very little sense of audience, limited genre or register awareness weak paragraphs, punctuation, etc.
1 Characteristics of "Fail"	Telegraphic, severely limited, may be rudimentary or unintelligible	Range limited to prompt repetition and low-level vocabulary – frequent errors in form and choice	Little to no ability to analyze or complete B2 task – very little production; repeats prompt	Little to no knowledge of conventions – limited to no audience awareness
0	—Too little to evaluate—			

APPENDIX B

English Penn Treebank Tag Set (Marcus et al., 1993, p. 317)

Tag	Description	Tag	Description
CC	Coordinating conjunction	TO	*to*
CD	Cardinal number	UH	Interjection
DT	Determiner	VB	Verb, base form
EX	Existential *there*	VBD	Verb, past tense
FW	Foreign word	VBG	Verb, gerund or present participle
IN	Preposition or subordinating conjunction	VBN	Verb, past participle
JJ	Adjective	VBP	Verb, non-3rd person singular present
JJR	Adjective, comparative	VBZ	Verb, 3rd person singular present
JJS	Adjective, superlative	WDT	Wh-determiner
LS	List item marker	WP	Wh-pronoun
MD	Modal	WP$	Possessive wh-pronoun
NN	Noun, singular or mass	WRB	Wh-adverb
NNS	Noun, plural	#	Pound sign
NNP	Proper noun, singular	$	Dollar sign
NNPS	Proper noun, plural	.	Sentence final punctuation (. ! ?)
PDT	Predeterminer	,	Comma
POS	Possessive ending	:	Mid-sentence punctuation (: ; - – …)
PRP	Personal pronoun	(	Left bracket character ({ [<
PRP$	Possessive pronoun	)	Right bracket character) }] >
RB	Adverb	‘	Left open single quote
RBR	Adverb, comparative	“	Left open double quote
RBS	Adverb, superlative	’	Right close single quote
RP	Particle	”	Right close double quote
SYM	Symbol		

APPENDIX C

Guidelines for Coding Errors (modified from Polio & Shea, 2014, pp. 24-25)

1. Whole clause is incomprehensible, intended structure is not clear, or there are more than five errors.
Example: And in the same time you might be sometime answered other people any questions.

2. Missing subject
Example: But sometime you might have you own secret that can't tell anybody except one person.

3. Missing verb (there is no verb in the clause)
Example: When he thinks he have to something, he does it finally even very difficult thing that other people give up.

4. Missing verb complement or object or required prepositional phrase
Example:.and I'm missing now.

5. Verb phrase problem: Wrong tense/aspect or malformed tense/aspect (including missing auxiliaries). Also wrong participle in a participle clause. Attempt at something passive-like where it does not belong.
Example: I have been studied there for eight months. It can be reduce the accident rate.

6. Preposition problem (missing, extra, wrong)
Example: And my brother-in-law graduated in MSU 10 year ago,

7. Sentence fragment
Example: I have five members. My parents (father, mother), younger sister, younger brother, and me.

8. Run-on sentence (Count the error in the first T-unit.)
Example: As time goes by and having more sense of being a part of this campus, I love to enjoy the great service provided by school such as gyms, libraries, labs in departmental building, all of them are well-organized and convenient for faculty and students to have a better living and do academic research.

9. Problem with relative clause formation including wrong relative pronoun, reduced relative clause (use of infinitive instead of participle), or resumptive pronoun
Example: It is the place that we enjoy in it.

10. Wrong modal or addition of modal where not needed
Example: Every day we can get many useful information from him.

11. Incorrect formation of passive voice including get passive (must be obviously passive)
Example: The building built by the construction company. (as opposed to something like: The building which is sat on the hill)

12. SV agreement
Example: She has the religion of buddism which mean she is a buddist.

13. Wrong pronoun or possessive determiner (including reflexive) and it/there.
Example: That's the reason why I don't like them. (them refers to father)

14. Quantifier–noun agreement (much/many, this/these) or other quantifier problems (a few/few); not including singular plural
Example: There are little students comparing MSU.

15. Problematic comparative or superlative formation
Example: In recent research, the capital of Korea, Seoul, is the worst clean city in the world.

16. Singular/plural error (including making mass nouns plural)
Example: Because there are all kinds of store around it.

17. Negation problem (including missing *do*, wrong word order related to negation)
Example: So, my father couldn't study no more

18. Wrong, extra, or missing article (for frequent English proper nouns, require appropriate article use but not for foreign words)
Example: From the middle of September to the end of November, it was a very nice scenery.

19. Wrong lexical item (including conjunctions, phrasal verb, formulaic chunks)
Example: Also, we have many green and colorful flowers in the yeard. [meaning on campus]

20. Wrong word form (e.g., adjective for noun) or wrong derivational formation.
Example: He has much patient. It is very crowdy.

21. Word order problem
Example: How did you stay for 13 hours every day in school?

22. Missing or extra word not included above
Example: I was really tired of routine work, stay late evening.

23. Severe punctuation error (not including run-on, don't include capitalization, be very lenient with comma errors) Include possessives such as "My brothers house" or contraction problems such as its/it's.
Example: That why, I have a time to do my work.

24. Gerund/infinitive (where the verb form should be either gerund or infinitive, or gerund and infinitive is misused)

Example: Make a good friend is difficult.
Example: I had such a great experience to study there.

25. Severe spelling error that causes a breakdown in meaning
Example: It's a great way to excacte your plan.

26. Others: e.g., genitive
Example: My university's friend (for my university friend)

APPENDIX D

Supplementary Statistical Analysis Results

Table D1

Kruskal-Wallis Test Results for Differences in Feature Frequencies Across Score Levels

Feature description	**Post hoc test results**	***p-value***
1. Post-nominal modification with		
1.1 Relative clause	Score 1 < Score 2 < Score 3 Score 1, Score 2 < Score 4 , Score 5	.024
2. *That*-complement clauses controlled by		
2.1 verb		.556
2.2 adjective		.662
2.3 noun		.304
3. *To*-complement clauses controlled by		
3.1 verb	Score 1 < Score 2, Score 3, Score 4	.017
3.2 verb (subject-to-object raising)	Score 1, Score 2 < Score 3, Score 4, Score 5	< .000
3.3 verb (passive)		.067
3.4 adjective		.435
3.5 adjective (subject-to-object raising)		.038
3.6 noun	Score 1, Score 2 < Score 3, Score 4, Score 5	< .000
4. *Wh*-word clauses (as subjects or objects)		
4.1 WH- *to*-complement		.266
4.2 WH-NP-VP	Score 1, Score 2 < Score 3, Score 4, Score 5	< .000
5. Ditransitive clauses		
5.1 Ditransitive (NP-V-NP-NP)		.064
5.2 Prepositional dative (NP-V-NP-PP)		.194

Table D2

Kruskal-Wallis Test Results for Differences in Component Scores Across Score Levels

Component	**Post hoc test results**	***p-value***
Component 1	Score 1 < Score 2 < Score 3 < Score 4, Score 5	< .001
Component 2	Score 1 < Score 2 < Score 3, Score 4	< .001
Component 3		.365
Component 4	Score 1, Score 2 < Score 3, Score 4, Score 5	< .001
Component 5		.819
Component 6	Score 1 < Score 2, Score 3, Score 4, Score 5 Score 2 < Score 4, Score 5	< .001

REFERENCES

REFERENCES

Alexopoulou, T., Michel, M., Murakami, A., & Meurers, D. (2017). Task effects of linguistic complexity and accuracy: A large-scale learner corpus analysis employing Natural Language Processing techniques. *Language Learning, 67*(S1), 180-208.

Arnold, T. B. (2018). cleanNLP: A tidy data model for Natural Language Processing. Retrieved from https://cran.r-project.org/web/packages/cleanNLP/cleanNLP.pdf

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

Banerjee, J., Franceschina, F., & Smith, A. M. (2007). Documenting features of written language production typical at different IELTS band score levels (IELTS Research Reports Volume 7). Retrieved from International English Language Testing System website: https://www.ielts.org/en-us/teaching-and-research/research-reports/volume-07-report-5

Banerjee, J., Yan, X., Chapman, M., & Elliott, H. (2015). Keeping up with the times: Revising and refreshing a rating scale. *Assessing Writing, 26,* 5-19.

Barker, F., Salamoura, A., & Saville, N. (2015). Learner corpora and language testing. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 511-534). Cambridge University Press.

Bestgen Y., & Granger, S. (2014). Quantifying the development of phraselological competence in L2 English writing: An automated approach. *Journal of Second Language Writing, 26,* 28-41.

Biber, D. (1988). *Variation across speech and writing*. Cambridge University Press.

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge University Press.

Biber, D., & Egbert, J. (2016). Register variation on the searchable web: A multi-dimensional analysis. *Journal of English Linguistics, 44*(2), 95-137.

Biber, D., & Gray, B. (2013). Discourse characteristics of writing and speaking task types on the TOEFL iBT test: A lexico-grammatical analysis (RR-13-04). Retrieved from https://doi.org/10.1002/j.2333-8504.2013.tb02311.x

Biber, D., & Gray, B., & Ponpoon, K. (2011). Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *TESOL Quarterly, 45*(1), 5-35.

Biber, D., Gray, B., & Staples, S. (2016). Predicting patterns of grammatical complexity across language exam task types and proficiency levels. *Applied Linguistics, 37*(5), 639-668.

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

Callies, M., Díez-Bedmar, M. B., & Zaytseva, E. (2014). Using learner corpora for testing and assessing L2 proficiency. In P. Leclercq, A. Edmonds & H. Hilton (Eds.), *Measuring L2 Proficiency: Perspectives from SLA* (pp. 71-90). Bristol: Multilingual Matters.

Chapelle, C. A. (2011). Validity argument for language assessment: The framework is simple…. *Language Testing, 29*(1), 19-27.

Chapelle, C. A. (2018, March). *Validity arguments in language assessment: Contributions from applied linguistics*. Plenary session presented at the Language Assessment Research Conference (LARC), Ames, IA, USA.

Chapelle, C. A., Enright, M. K., & Jamieson, J. (2008). *Building a validity argument for the Test of English as a Foreign Language*. London: Routledge.

Chapelle, C. A., Enright, M. K., & Jamieson, J. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice, 29*(1), 3-13.

Chapelle, C. A. (March, 2018). Validity argument in language assessment: Contributions from Applied Linguistics. Opening address presented at the Language Assessment Research Colloquium, Ames, IA.

Council of Europe. (2001). *The Common European Framework of Reference for languages: Learning, teaching, assessment*. Retrieved from https://www.coe.int/t/dg4/linguistic/Source/Framework_EN.pdf

Council of Europe. (2018). Common European Framework of Reference for languages: Learning, teaching, assessment: Companion volume with new descriptors. Retrieved from https://rm.coe.int/cefr-companion-volume-with-new-descriptors-2018/1680787989

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., Baba, K., Eouanzoui, K., Erdosy, U., & James, M. (2005). Analysis of discourse features and verification of scoring levels for independent and integrated prototype written tasks for the new TOEFL (RR-05-13, TOEFL-MS-30). Retrieved from http://dx.doi.org/10.1002/j.2333-8504.2005.tb01990.x

Cushing, S. T. (2017). Corpus linguistics in language testing research. *Language Testing, 34*(4), 441-449.

Deygers, B., Zeidler, B., Vilcu, D., & Carlsen, C. H. (2017). One framework to unite them all? Use of the CEFR in European university entrance policies, *Language Assessment Quarterly*, doi: 10.1080/15434303.2016.1261350

Egbert, J. (2017). Corpus linguistics and language testing: Navigating uncharted waters. *Language Testing, 34*(4), 555-564.

Egbert, J. & Biber, D. (2018). Do all roads lead to Rome?: Modeling register variation with factor analysis and discriminant analysis. *Corpus Linguistics and Linguistic Theory, 14*(2), 233-273.

English Profile. (2015). English Grammar Profile. Retrieved from http://englishprofile.org/english-grammar-profile

Evans, N. W., Hartshorn, K. J., Cox, T. L., & Martin de Jel, T. (2014). Measuring written linguistic accuracy with weighted clause ratios: A question of validity. *Journal of Second Language Writing, 24*, 33-50.

Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. London: SAGE.

Friginal, E., & Weigle, S. C. (2014). Exploring multiple profiles of L2 writing using multidimensional analysis. *Journal of Second Language Writing, 26*, 80–95.

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

Gries, S. Th. (2013). *Statistics for linguistics with R* (Second edition). DeGruyter Mouton.

Hawkins, J. A., & Buttery, P. (2010). Criterial Features in learner corpora: Theory and illustrations. *English Profile Journal, 1*, 1-23.

Hawkins, J. A., & Filipović, L. (2012). *Criterial features in L2 English: Specifying the reference levels of the Common European Framework*. Cambridge: Cambridge University Press.

Hripcsak, G., & Rothschild, A. S. (2005). Agreement, the *f*-measure, and reliability in information retrieval. *Journal of the American Medical Informatics Association, 12*(3), 296-298.

Hunt, K. W. (1965). *Grammatical structures written at three grade levels* (Research Report No. 3). Champaign, IL: National Council of Teachers of English. Retrieved from https://eric.ed.gov/?id=ED113735

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2014). *An introduction to statistical learning: With applications in R*. Springer Publishing Company, Incorporated.

Jarvis, S., Grant, L., Bikowski, D., & Ferris, D. R. (2003). Exploring multiple profiles of highly rated learner compositions. *Journal of Second Language Writing, 12*, 377-403.

Johnson, M. D. (2017). Cognitive task complexity and L2 written syntactic complexity, accuracy, lexical complexity, and fluency: A research synthesis and meta-analysis. *Journal of Second Language Writing, 37*, 13-38.

Juknevičienė, R., & Šeškauskienė, I. (2014). The National Examination of English in Lithuania: Searching for evidence of CEFR criterial achievement levels. *Studies about Languages, 25*, 88-96.

Kane, M. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*(3), 527-535.

Kane, M. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17-64). Westport, CT: American Council on Education and Praeger.

Kane, M. (2011). Validating score interpretation and uses: Messick Lecture, Language Testing Research Colloquium, Cambridge, April 2010. *Language Testing, 29*(1), 3-17.

Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1-73.

Knoch, U., & Chapelle, C. A. (2018). Validation of rating processes within an argument-based framework. *Language Testing, 35*(4), 477-499.

Knoch, U., Macqueen, S., & O'Hagan, S. (2014). An investigation of the effect of task type on the discourse produced by students at various score levels in the TOEFL iBT writing test (RR-14-43). Retrieved from http://dx.doi.org/10.1002/ets2.12038

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

LaFlair, G., & Staples, S. (2017). Using corpus linguistics to examine the extrapolation inference in the validity argument for a high-stakes speaking assessment. *Language Testing, 34*(4), 451-475.

Lahuerta Martínez, A. C. (2018). Analysis of syntactic complexity in secondary education ELF writers at different proficiency levels. *Assessing Writing, 35*, 1-11.

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.

Manning, C. D. (2011). Part-of-Speech tagging from 97% to 100%: Is it time for some linguistics? In A. F. Gelbukh (Ed.), *Computational Linguistics and Intelligent Text Processing* (pp. 171-189). Berlin, Heidelberg: Springer.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky. D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 55-60).

Marcus, M. P., Santorini, B., Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics, 19*(2), 313-330.

Matthews, J., & Wijeyewardene, I. (2018). Exploring relationships between automated and human evaluations of L2 texts. *Language Learning & Technology, 22*(3), 143-158.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-104). New York, NY: American Council on education and Macmillan.

Michigan State University English Language Examinations. (n.d.). MSU Exams. Retrieved from http://www.msu-exams.gr/en

O'Keeffe, A., & Mark, G. (2017). The English Grammar Profile of learner competence. *International Journal of Corpus Linguistics, 22*(4), 457-489.

Park, J.-H. (2017). *Syntactic complexity as a predictor of second language writing proficiency and writing quality*. Unpublished doctoral dissertation, Michigan State University, East Lansing, Michigan.

Park, K. (2014). Corpora and language assessment: The state of the art. *Language Assessment Quarterly, 11*, 27-44.

Polio, C. (1997). Measures of linguistic accuracy in second language writing research. *Language Learning, 47*(1), 101-143.

Polio, C., & Shea, M. (2014). An investigation into current measures of linguistic accuracy in second language writing research. *Journal of Second Language Writing, 26*, 10-27.

R Core Team. (2019). R: A language and environment for statistical computing [Computer software]. R Foundation for Statistical Computing, Vienna, Austria.

Römer, U. (2009). The inseparability of lexis and grammar: Corpus linguistic perspectives. *Annual Review of Cognitive Linguistics, 7*(1), 140-162.

Sireci, S. G. (2009). Packing and unpacking sources of validity evidence: History repeats itself again. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions, and applications* (pp. 19-37). Charlotte, NC: IAP Information Age Publishing.

Staples, S., Egbert, J., Biber, D., & Gray, B. (2016). Academic writing development at the university level: Phrasal and clausal complexity across level of study, discipline, and genre. *Written Communication, 33*(2), 149-183.

Staples, S., LaFlair, G. T., & Egbert, J. (2017). Comparing language use in oral proficiency interviews to target domains: Conversational, academic, and professional discourse. The *Modern Language Journal, 101*(1), 194-213.

Staples, S., & Reppen, (2016). Understanding first-year L2 writing: A lexico-grammatical analysis across L1s, genres, and language ratings. *Journal of Second Language Writing,*

32, 17-35.

Straka, M., & Straková, J. (2017). Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, Vancouver, Canada, August 2017. Retrieved from http://ufal.mff.cuni.cz/~straka/papers/2017-conll_udpipe.pdf

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th edition). Pearson.

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

Taylor, L., & Barker, F. (2008). Using corpora for language assessment. In N. H. Hornberger (Ed.), *Encyclopedia of language and education* (2008 ed.). Retrieved from https://doi.org/10.1007/978-0-387-30424-3_179

Thewissen, J. (2013). Capturing L2 accuracy developmental patterns: Insights from an error-tagged EFL learner corpus. *Modern Language Journal, 97*(S1), 77-101.

Timm, J. (2018). corpuslingr: Some corpus linguistics in R. Retrieved from https://github.com/jaytimm/corpuslingr

Toulmin, S. E. (2003). *The uses of argument*. Cambridge, UK: Cambridge University Press.

Verspoor, M., Schmid, M., & Xu, X. (2012). A dynamic usage based perspective on L2 writing. *Journal of Second Language Writing, 21*, 239-263.

Wisniewski, K. (2017). Empirical learner language and the levels of the Common European Framework of Reference. *Language Learning, 67*(S1), 232-253.

Wisniewski, K. (2018). The empirical validity of the Common European Framework of Reference Scales: An exemplary study for the vocabulary and fluency scales in a language testing context. *Applied Linguistics, 39*(6), 933-959.

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

Xi, X. (2017). What does corpus linguistics have to offer to language assessment? *Language Testing, 34*(4), 565-577.

Yan, X., & Staples, S. (2016). Investigating lexico-grammatical complexity as construct validity evidence for the ECPE writing tasks: A multidimensional analysis (CaMLA Working Papers 2016-01). Retrieved from Cambridge Michigan Language Assessment website: http://cambridgemichigan.org/wp-content/uploads/2017/05/CWP-2016-01-Yan-Staples.pdf