CaMLA Working Papers

2016–01

Investigating Lexico-grammatical Complexity as Construct Validity Evidence for the ECPE Writing Tasks: A Multidimensional Analysis

Xun Yan
Shelley Staples

Investigating Lexico-grammatical Complexity as Construct Validity Evidence for the ECPE Writing Tasks: A Multidimensional Analysis

Authors

Xun Yan
Shelley Staples

About the Authors

Xun Yan is Assistant Professor of Linguistics, SLATE, and Educational Psychology at University of Illinois at Urbana-Champaign (UIUC). He is also the supervisor of the English Placement Test (EPT) at UIUC. His research interests include quality management of post-admission language assessments, assessment literacy for language teachers, formulaic language acquisition, and test score use in educational settings. His work has been published in *Language Testing, Assessing Writing*, and *Journal of Second Language Writing*.

Shelley Staples is Assistant Professor of English (English Applied Linguistics and SLAT) at University of Arizona. Her research focuses on corpus analyses of specialized spoken and written registers, particularly for applications to L2 writing and assessment. Her work has recently been published in journals such as *Applied Linguistics*, *Journal of English for Academic Purposes*, and *Journal of Second Language Writing*.

Table of Contents

Abstract

The complexity of lexico-grammatical features is widely recognized as an integral part of writing proficiency in second language (L2) writing assessment. However, a remaining concern for the construct validation of writing tasks lies in the scalability of representative linguistic features in writing performances. Previous research suggests that distinctions across different levels of writing proficiency are not necessarily associated with individual lexico-grammatical features, but rather with the co-occurrence of multiple features (Biber, Gray, & Staples, 2016; Friginal, Li & Weigle, 2014; Jarvis, Grant, Bikowski & Ferris, 2003).

As an effort to investigate the scalability of lexico-grammatical complexity, this study used a multidimensional (MD) analysis to examine saliency and patterns of co-occurrence for 31 lexico-grammatical features in 595 writing performances on a large-scale, advanced-level English language proficiency examination, the Examination for the Certificate of Proficiency in English (ECPE). The linguistic features were classified into four categories: fluency, lexical sophistication, semantic categories for word classes, and general grammatical features, all of which have been found to characterize written discourse and advanced L2 writing proficiency (e.g., Biber, Gray, & Staples, 2016).

Results of the MD analysis indicate five underlying factors, representing five functional dimensions of lexico-grammatical complexity in ECPE writing performances: literate vs. oral discourse, topic-related content, prompt dependence vs. lexical diversity, overt suggestions, and stance vs. referential discourse. Together, the five dimensions accounted for 35% of the holistic score variance. While factor scores on the prompt-difference dimension did not yield significant correlation with the holistic ECPE writing scores awarded by human raters, correlations for the other four dimensions were linear and statistically significant. Among these four dimensions, only three dimensions demonstrated significant differences across essays of different score levels. Findings of this study present supportive evidence for different shades/layers of construct validity of ECPE writing tasks and suggest the scalability of the ECPE writing scale with respect to lexico-grammatical complexity.

Introduction

The complexity of lexico-grammatical features is widely recognized as an integral part of writing proficiency in second language (L2) writing assessment. However, a remaining concern for the construct validation of writing tasks lies in the scalability of representative linguistic features in writing performances. Previous research suggests that distinctions across different levels of writing proficiency are not necessarily associated with individual lexico-grammatical features, but rather with the co-occurrence of multiple features (Biber, Gray, & Staples, 2016; Friginal, Li & Weigle, 2014; Jarvis, Grant, Bikowski, & Ferris, 2003). This study further investigates lexico-grammatical features in a representative ECPE essay corpus as construct-related validity evidence for the ECPE writing section using multidimensional (MD) analysis. Specifically, we identified patterns of co-occurrence among lexico-grammatical features that help distinguish ECPE writing performances across score levels. Findings of this study support the scalability of lexico-grammatical features on the ECPE writing scale. More importantly, relationships between different dimensions of lexico-grammatical complexity and holistic essay scores provided supportive evidence for the construct-related validity of the ECPE writing tasks.

Literature review

Scalability of lexico-grammatical complexity as construct validity for writing assessment

The complexity of lexico-grammatical features has been recognized as a core component of second language (L2) writing proficiency and is often used as an

effective indicator of stages of L2 writing development and differences in L2 writing proficiency (Biber et al., 2016; Lu, 2011). In L2 writing assessment, lexico-grammatical complexity has been frequently examined as a key component of the construct of writing proficiency (Weigle, 2002). The linguistic features associated with L2 writing proficiency include among others: lexical profile (e.g., range), formulaic language (e.g., lexical bundles), the use of a range of grammatical categories (e.g., nominalizations and adjectives), and complex phrasal (e.g., noun phrase construction) and clausal structures (e.g., subordination) (see Biber & Gray, 2013, for a summary of studies investigating lexico-grammatical features in spoken and written discourse). The use of both simple and complex syntactic and lexical features, especially in timed writing situations, is believed to represent an integral part of a writer's language proficiency and to facilitate the formulation of meaning and relationships at multiple layers in text.

The ECPE writing section includes lexical and grammatical features as two of the major analytic components in its rating scale. Despite the essential role of lexico-grammatical complexity, an important concern for the validation of writing assessments remains in the scalability of the the construct of lexico-grammatical complexity. Specifically, it is unclear whether the linguistic features incorporated in the rating scale can truly distinguish performances across score levels. As suggested by Ortega (2003) in her meta-analysis of syntactic complexity measures, many lexico-grammatical features might have a curvilinear relationship with overall writing proficiency. In addition, Biber et al. (2016) found few relationships between individual features and score level on the TOEFL iBT; however, when an array of linguistic features was reduced to fewer underlying factors, the factor scores reflected stronger distinctions and a linear trend across score levels. This finding was partially corroborated by a previous scale revision study on ECPE (Banerjee, Yan, Chapman, & Elliot, 2015), where syntactic complexity, operationalized as a single holistic measure (i.e., number of modifiers per noun in Coh-Metrix), did not yield a consistent linear increase among ECPE essays as the holistic score level increased. These findings suggest that, on the one hand, individual linguistic features may not reliably distinguish writing performance across score levels; however, on the other hand, the co-occurrence of multiple lexico-grammatical features is arguably a better approach to distinguishing writing performance across proficiency levels, thus further strengthening the argument for the construct validity of the test (Biber et al., 2016; Friginal et al., 2014; Jarvis et al., 2003).

Multidimensional analysis of lexico-grammatical complexity in writing assessment

A method to investigate co-occurrence patterns of lexico-grammatical features is multidimensional (MD) analysis, a corpus-based analytic framework developed by Biber (1988) for exploring linguistic variations in spoken and written English texts. MD analysis, derived from corpus linguistic techniques and factor analysis, accounts for the co-variation among a wide array of lexico-grammatical features and reduces these features to a smaller number of functionally interpretable linguistic dimensions. The advantage of MD analysis is that it represents writing proficiency through a few holistic linguistic dimensions while accounting for all the individual features that contribute to the linguistic dimensions in the analysis of writing performance. Although MD analysis has been applied to studying L2 writing in a number of studies (e.g., Biber et al., 2016; Cao & Xiao, 2013; Weigle & Friginal, 2015), this method has not been used for analyzing the performance on the ECPE writing task.

Multidimensional analysis has primarily focused on grammatical features, such as personal pronouns, and syntactic structures, such as verb and noun complement clauses. However, there are certain exceptions to this, namely in the form of type/token ratio and word length, which have been included in MD analyses since Biber's initial (1988) study. More recently, semantic categories of adjectives, nouns and verbs have also been added. Biber, Gray and Staples (2016), for example, incorporated semantic categories of nouns and adjectives into their multidimensional analysis of TOEFL iBT essays. Egbert (2015), in his study of professional academic writing, includes both frequency of core and academic vocabulary as well as commonly used lexical bundles in his MD analysis. Finally, Staples, LaFlair, and Egbert (2014) investigated the impact of vocabulary frequency alongside lexico-grammatical features in an MD analysis of the MELAB speaking assessment. We follow these more recent studies to incorporate both grammatical and lexical aspects of language use in our MD analysis.

In particular, vocabulary frequency has been studied as a predictor of development, with lower proficiency learners using more high frequency words and higher proficiency learners using fewer high frequency words and more low frequency words (e.g., Laufer & Nation,

1995). Another well-studied domain of lexical use is lexical bundles. The popularity of research on lexical bundles originates from its psycholinguistic properties, i.e., holistic storage and access. The processing advantages of lexical bundles or formulaic language in general point to the largely lexical nature of lexical bundles, despite the variation in the syntactic distribution of lexical bundles. However, scholars have shown the existence of different syntactic and functional purposes among lexical bundles (e.g., Biber, Conrad, & Cortes, 2004; Simpson-Vlach & Ellis, 2010). For example, Biber and his colleagues (2004) classified lexical bundles into different syntactic categories, e.g., noun phrase fragment, preposition phrase fragment, verb phrase fragment. Additionally, in both Biber et al. (2004) and Simpson-Vlach and Ellis (2010), lexical bundles have been classified into a number of functional categories, e.g., referential expressions, stance expressions, and discourse organizers. The emergence of syntactic and functional variations of lexical bundles suggests that both written and spoken discourse is formulaic to a certain extent. More importantly, writers or speakers use lexical bundles to facilitate the formulation of speech and writing and to perform different functions. Therefore, we argue that the use of lexical bundles should be included as a feature of lexico-grammatical complexity.

Research Questions

This study investigated the lexico-grammatical features of ECPE writing performance through an MD approach. It also examined relationships between individual and co-occurring linguistic features and the different levels of the ECPE writing scale. Specifically, the study seeks to address the following research questions:

1. What linguistic features in test-taker performances are associated with different levels of the ECPE writing scale?
2. In what ways are test scores associated with systematic linguistic differences in test-taker performance on the ECPE writing tasks?
3. Do test takers systematically vary the linguistic features of writing performance in response to different ECPE writing prompts? If so, how?

Methods

ECPE corpus

The corpus used for this study comprised 595 essays from the ECPE writing section. The essays were identified through stratified random sampling to represent the distribution of writing proficiency level (see Table 1; A-E represent the holistic score bands for the ECPE, with A as the highest level) and examinee L1 background on the test and were evenly drawn from three prompts (see Table 1). A total of 600 essays were collected; however, five essays were excluded from the final corpus because they were too short (i.e., less than 100 words) to generate reliable analysis results. The final version of the ECPE corpus consisted of 177,163 words, where essay length ranged from 104 to 513 words, with an average length of around 300 words (*M = 295.27, SD = 70.73*). The overwhelming majority of examinees were from Southern Europe and South America, where the ECPE is widely used for academic and professional purposes. Three L1s were most represented: Greek (88.7%), Portuguese (3.5%), and Spanish (4.7%). There were more female examinees (59%) than male examinees (41%). Age of the examinees ranged from 13 to 67 years, though the majority of the examinees were around the age of 20 (M=20.93, SD=7.94). Table 2 shows the topics of the three prompts represented in the ECPE corpus. Each essay was transcribed in a computer readable format by a trained transcriber.

Table 1: The ECPE essay corpus

	A	B	C	D	E	Total
Prompt 1	30	47	60	47	15	**199**
Prompt 2	30	41	60	41	25	**197**
Prompt 3	30	46	60	47	16	**199**
Total	**90**	**134**	**180**	**135**	**56**	**595**

Table 2: Writing prompts represented in the ECPE corpus

Prompt	Topic
P1	Use of text messages and internet chat rooms to communicate among teenagers
P2	Suggestion on the use of educational technologies in the classroom
P3	Advantages and disadvantages of a country having a large tourism industry

Variables included in the study

There was an array of lexico-grammatical features investigated in this study, all chosen based on previous MD analyses of writing (e.g., Biber, 1988; Biber et al., 2016; Egbert, 2015). These features include grammatical forms (e.g., attributive adjectives), syntactic structures (e.g., finite adverbial clauses), semantic categories of grammatical forms (e.g., communication verbs, activity verbs), vocabulary features (frequency, type/token ratio, word length) and lexical bundles. The variables were subjected to a multidimensional analysis (see Data Analysis for detailed discussion of the method) to extract a smaller number of linguistic and functional dimensions underlying the individual features. Table 3 displays the individual features included in the multidimensional analysis. The three categories of features were operationalized through different automatic tools, which are further described below.

Grammatical and syntactic features. Most of the features were annotated automatically using the Biber tagger and counted using a program (also developed by Biber) called TagCount (Biber, 1988, 2006). The Biber tagger is a rule-based and probability-based tagger that has been widely used in corpus research since the late-1980s. TagCount is a program that automatically calculates the normed rates of occurrence (per 1,000 words) for more than 150 linguistic features in corpus texts. However, in this study we investigate the use of only a small subset of these linguistic variables. We conducted an analysis of accuracy (precision and recall) on 10% of the data for variables previously identified as problematic (see Biber & Gray, 2013, pp. 16-18). After this analysis, it was deemed necessary to correct the tags for several features. Some of these features were corrected automatically using Perl scripts. Other features (noun + *that* complement clauses, relative clauses, and present

Table 3: Variables included in the multidimensional analysis

Variable	Description/example
Grammatical and syntactic features	
Finite adverbial clauses	*…**because** they are "English" in their customs and practices.*
Likelihood verb + that clause	We **think *that*** *Eulalie writes or speaks her monologue…*
Stance noun + that clauses	The **belief *that*** *people could be distinguished by…*
Present tense verbs	verb (uninflected present, imperative & third person)
Have as a main verb	
Mental verbs	*know, think, believe*
Certainty verbs	*conclude, prove, show, understand, find, know, realize*
Communication verbs	*argue, claim, propose, say, tell, suggest*
Activity verbs	*smile, open*
Attitude verbs	*anticipate, expect, prefer*
Passive voice	Racism ***was rejected*** as a scientific concept.
Necessity modals	*should, must, have to*
Pronoun *it*	
Third person pronouns	*they, she, he*
Attributive adjectives	***industrial*** scale, ***social*** reality
Size adjectives	*large, big, high, long*
Time adjectives	*new, young, old*
Topical adjectives	*political, international, national, economic*
Premodifying nouns	***metal*** ions, ***crime*** rate
Common nouns	*inequalities, ontology, formula, proteins*
Cognition nouns	*fact, knowledge*
Place nouns	*country, city, continent, border, mountain, ocean*
Human nouns	*family, parent, teacher, official, president, people*

Table 3: Variables included in the multidimensional analysis

Variable	Description/example
Group nouns	*committee, congress, bank*
Process nouns	*application, meeting, balance*
Technical nouns	*internet, web*
Concrete nouns	*computer, machine, equipment, video*
Nominalizations	*collectivists, existence, sweetness*
Indefinite articles	*a, an*
Definite articles	*the*
Prepositions	*in, at, on*
Lexical bundles	
Proportion of prompt-match bundles	*with their friends, a large tourism industry, to communicate with*
Proportion of generic lexical bundles	*in order to, on the other hand, a lot of*
Proportion of stance bundles	*my opinion is, they should not, it is very important*
Proportion referential bundles	*a lot of, more and more, the opportunity to*
Vocabulary features	
Word count	Total number of words
Word length	Average word length
Vocabulary not in the first 3000	Vocabulary less frequent than the first 3000 words
Vocabulary 501- 3000	Vocabulary in the 501-3000 word frequency range
Vocabulary 1-500	Vocabulary in the 1-500 frequency range
Type token ratio	Ratio of types to tokens in the first 400 words of each text

participles) were corrected manually in the corpus using an interactive fix-tagging program developed for use in Biber and Gray (2013). Semantic categories for a number of the grammatical forms were also included in the data analysis. These categories acknowledge the fact that, in usage-based approaches to development, lexis and grammar are learned in concert rather than individually.

Vocabulary features. Type/token ratio (TTR) and word length were calculated through the Biber Tagger. The TTR is limited to the first 400 words in each essay, in order to account for variation in TTR based on word count. This means that TTR for longer essays was calculated using the first 400 words, whereas TTR for shorter essays (shorter than 400 words) was calculated using the entire essay. We measured word frequency using an online tool called WordandPhrase (Davies, 2011). This tool uses a large list of the most frequent words in English, based on the COCA corpus, to report the percent of a text that is composed of words in various frequency bands, 1-500 most frequent words, 501-3,000 most frequent words, and words that are not among the 3,000 most frequent words.

Lexical bundles. Lexical bundles examined in the ECPE essay corpus were three- to five-word bundles extracted from the ECPE corpus in a bottom-up approach using a programing script written in the R language (Banerjee et al., 2015). The extracted bundles were first classified into three subcategories: prompt-match bundles (i.e., bundles that are exact matches with the prompt), topic-related bundles (i.e., bundles that are not exact matches but related to the topic of the prompt), and generic bundles (i.e., bundles that are not related to the content of the prompt). This resulted in 289 lexical bundles, which included 29 prompt-match lexical bundles, 90 topic-related lexical bundles, and 170 generic lexical bundles. Among the 29 prompt-match bundles, 11 bundles were exact matches for prompt 1, 7 for prompt 2, and 11 for prompt 3. The 170 generic lexical bundles were further classified into three functional types, to further investigate the use of different types of lexical bundles by writers across

proficiency levels (adapted from Biber et al., 2004; Simpson-Vlach & Ellis, 2010; see the Appendix for the coding scheme).

To analyze the use of lexical bundles in individual essays, a second R programing script was written to automatically compute the total number of bundles and the frequency of each bundle in each essay. Individual bundle frequencies were then used to calculate the proportion of bundles for each subcategory (i.e., prompt-match, topic-related, generic). Within generic bundles, proportions for each functional type were also calculated. These transformation procedures resulted in five proportion-related variables for lexical bundle use.

Multidimensional analysis

The first step in our MD analysis was to carefully select 41 linguistic features to include in the factor analysis (see Table 3). The next step was to perform a factor analysis on the normed rates of occurrence for the full set of 41 linguistic features to determine whether they could be reduced to a smaller set of interpretable dimensions. The statistical software R was used to perform this exploratory factor analysis (R Core Team, 2015). We used the R function 'fa' (factor analysis), which is part of the 'psych' library, using principal axis factoring and a Promax rotation. The Kaiser-Meyer-Olkin Measure of Sampling Adequacy (KMO) was .73, which is a 'middling' but still acceptable number for continuing with the factor analysis. The scree plot of eigenvalues revealed a definitive break between the fifth and sixth factor. Therefore, a five factor solution was chosen. The cumulative variance accounted for by the five factors was 35%. Variables were only included in the analysis if they met a minimal factor loading threshold of +/-.30. After assigning each of the variables to the factor where it loaded the strongest, the positive-loading features were separated from the negative-loading features.

After we established the factors where each variable loaded, we calculated dimension scores for each of the 595 texts in the ECPE corpus. This was done in two steps. First, we standardized the rates of occurrence for each linguistic feature to a mean of 0 and a standard deviation of 1 using the z-score formula. Second, we summed the standardized counts for the negative loading features and subtracted them (where applicable) from the sum of the counts for the standardized positive-loading features for each dimension. This resulted in five dimension scores for each text.

The final step in the MD analysis was to explore the underlying functional interpretation of each factor and assign a dimension label to each of the factors. We relied on two sources of information to complete this step for each of the five dimensions: (a) the co-occurrence patterns for the linguistic features, and (b) the use of these linguistic features in the texts. Explanations for the dimension labels we chose are included in the results section below.

Data analysis

The dimension scores for each of the five dimensions were used to compare the ECPE essays across score level and prompt. We first conducted correlation analyses to determine whether there were significant relationships between score level and the five dimension scores. Then, we conducted a two-way factorial MANOVA on a subsample of 296 ECPE essays, which contained an approximately equal number of essays across prompts and ECPE score levels. A subsample was drawn with an equal number of observations in each cell to ensure that the statistical assumptions for MANOVA would be satisfied. The factorial MANOVA included score level (with 5 levels) and prompt (with 3 levels) as between-subject factors and the interaction effect between prompt and score level. Subsequently, post hoc univariate ANOVAs were also performed on all of the five dimensions to examine whether and how individual dimensions differed across score levels. Because there were five dimensions, the significance level was adjusted to .01 for the univariate ANOVAs, while a significance level of .05 was used for the factorial MANOVA and correlational analyses. Overall, these analyses allowed us to determine (1) whether the constellation of features on different dimensions was significantly different across ECPE score levels, and (2) whether there were any prompt differences reflected in the linguistic dimensions among ECPE essays across score levels.

Results and Discussion

Functional dimensions from MD analysis

We first present the results of the factor analysis, providing details about the five functional dimensions that characterize the ECPE writing performance. We explain our interpretations based on the individual linguistic features loading on each dimension and

illustrate the type of discourse that is characterized by each dimension.

The five dimensions are as follows:

Dimension 1: Literate vs. oral discourse

Dimension 2: Topic-related content

Dimension 3: Prompt dependence vs. lexical variety

Dimension 4: Overt suggestions

Dimension 5: Stance vs. referential discourse

Dimension 1: Literate vs. oral discourse. Dimension 1 in our study is similar to Dimension 1 in many other multidimensional analyses (e.g., Biber, 1988; Biber, 2006; Biber et al., 2016), with linguistic features that are associated with written language on one end of the spectrum (word length, attributive adjectives, passive voice, vocabulary with frequencies less than 500 words). The other end of the spectrum contains linguistic features more associated with spoken language (mental verbs, *have* as a main verb, subordinating conjunctions, and vocabulary with frequencies between 1-500).

Table 4: Linguistic features loading on Dimension 1, Literate vs. oral discourse

Feature	Factor Loading
Positive loadings	
Word length	0.82
Vocabulary 501-3000	0.69
Vocabulary not in the first 3000	0.67
Attributive adjectives	0.57
Passive voice	0.40
Negative loadings	
Vocabulary 1-500	-0.86
Finite adverbial clauses	-0.32
Mental verbs	-0.31
Have as a main verb	-0.31

Excerpt 1 provides an example of the literate discourse by showing more features with positive loadings on Dimension 1, while Excerpt 2 provides an example of essays showing more features with negative loadings (oral discourse).

Excerpt 1 File A_P2_69 Dimension 1 score = +3.89

Note: Attributive adjectives are bolded; passive voice is underlined; vocabulary frequency 501-3000 is in small caps and vocabulary not in the first 3000 is in large caps.

It is an **UNDENIABLE** fact of MODERN, **FUTURISTIC** SOCIETIES that students can (and are) PROFOUNDLY INFLUENCED by the CORNUCOPIA of **TECHNOLOGICAL** INNOVATIONS. As a CONSEQUENCE, the MAJORITY of them make use of **TECHNOLOGICAL** DEVICES such as CALCULATORS, COMPUTERS, ETC., in class. … Lastly, ASIDE from the BENEFTIS INSIDE the school, TECHNOLOGY EDUCATED students will also find it EASIER to COPE with EVERYDAY life, which of course is DOMINATED by TECHNOLOGY.

Excerpt 2 File D_P2_21 Dimension 1 score = -15.87

Note: Finite adverbial clauses are in bold, mental verbs are underlined, and vocabulary 1-500 word frequency is in small caps.

ALTHOUGH text messing IS GOOD WHEN YOU WANT TO communicate WITH XX THAT IS miles AWAY OR IN THE OTHER COUNTRY. I THINK THAT THE PARENTS SHOULD DO SOMETHING XX TELL THEIR CHILDREN NOT TO USE THEIR **cellphone** AND GET OUT MEET THEIR FRIENDS AND DO SOMETHING ELSE. … **I HOPE** THAT THEIR CHILDREN WILL UNDERSTAND AND STOP USING **text messaging** ALL DAY LONG.

As can be seen, the first excerpt, an example of more literate discourse, contains much more sophisticated language (both in terms of less frequent vocabulary and longer words). The writer also uses more attributive adjectives and one instance of passive voice (*which of course is dominated by technology*). On the other hand, the second excerpt uses more language associated with speech (shorter words, higher frequency words, and mental verbs like *hope* and *think*). There are also instances of finite adverbial clauses, which have been

shown to be used more in speech than in writing (Biber, Gray, & Poonpon, 2011).

Dimension 2: Topic-related content. Dimension 2 is particular to the ECPE, and reveals the impact of the topics within each prompt (this will be discussed further below along with the ANOVAs across prompts). Two main topics that this Dimension reveals are those related to technology and the future (particularly concrete and technical nouns, and time adjectives) and those related to place (place nouns). In our data, Prompt 1 is most specifically related to technology and 3 related to place (tourism). Prompt 2 has elements of both the technological elements (positive loadings) and place (negative loadings) since it focuses on the use of technology in schools.

Table 5: Linguistic features loading on Dimension 2, Topic-related content

Feature	Factor Loading
Positive loadings	
Technical nouns	0.65
Concrete nouns	0.54
Third person pronouns	0.45
Time adjectives	0.44
Attitude verbs	0.43
Communication verbs	0.38
Negative loadings	
Place nouns	-0.76
Nominalizations	-0.64
Size adjectives	-0.48
Definite articles	-0.32

Excerpt 3 File: C_905272768 Dimension 2 score = +14.40

Note: Bold indicates technical and concrete nouns, italics time adjectives, and third person pronouns are in small caps.

Firstly, **communicating** with other people or friends through **Internet** chat **rooms** and **text** messages is believed by *young* people to be something useful because THEY learn things. Furthermore THEY do also have face-to-face communication because *young* people are going out and meet THEIR friends so THEY are having face-to-face communication. But many *young* people do not hang out with THEIR friends and THEY use **computer** and cellular **phones** all day and night.

This excerpt illustrates the use of technical and concrete verbs in the responses to prompts that focus on new technology. The prompts focusing on technology also emphasize differences across age groups, thus eliciting more time adjectives (e.g., *young*).

Excerpt 4 File: C_907339865 Dimension 2 score = -11.85

Note: Bold indicates place nouns, italics size adjectives, underlining definite articles, and nominalizations are in small caps.

Moreover, people from the **country** have the OPPORTUNITY to learn new **cultures** and meet people from all over the **world**. Another *big* fact that TOURISM helps is the economy. It really is essential for a **country** to have a *large* TOURISM industry because, that way, the national economy will be improved at all points.

Excerpt 4 is a response to the topic of travel and tourism. Here, we can see that the writer uses more place nouns (*country, world*) and nominalizations (particularly *tourism*, due to the use of this word in the prompt). The greater use of definite articles is based on the use of nouns such as *economy*, and *world*, which are related to the topic of tourism.

We can see from these examples how much of a role the topic of the prompt plays in the use of language identified for this dimension. The categories of nouns (e.g., technical vs. place, nominalizations) and adjectives (e.g., time vs. size adjectives) in particular are greatly influenced by whether the prompt asks the test takers to write about technology or tourism.

Dimension 3: Prompt dependence vs. lexical variety. Like Dimensions 1 and 2, Dimension 3 contains two poles, one associated with prompt dependence (prompt-match bundles, common nouns, and premodifying nouns) and the other associated with lexical variety (generic bundles, type/token ratio, and the pronoun *it*). These features can be seen in the two excerpts below, one containing a great deal of prompt repetition, and the other containing a wider variety of vocabulary.

Table 6: Linguistic features loading on Dimension 3, Prompt dependence vs. lexical variety

Feature	Factor Loading
Positive loadings	
Common nouns	0.67
Proportion of prompt bundles	0.61
Premodifying nouns	0.54
Negative loadings	
Proportion of generic bundles	-0.58
Type/token ratio	-0.48
Pronoun *it*	-0.32

Excerpt 5 File E_P1_81 Dimension 3 score = +18.73

Note: Prompt bundles are in bold and premodifying nouns are italicized.

More specifically, **young people increasingly use *text* messaging and *internet chat* rooms to communicate with** friends. This situation **many times has result of many parents are worried that their children are not developing the skill of speaking to people face to face**. … The increase of use ***text* messaging** and ***internet chat* rooms** of young and generally all the age of people to communicate with theirs friends, can worried parents because are feeling that children are losing the connection with other children, more specifically connection about **face to face**.

As can be seen from Excerpt 5, the writer uses a great deal of language from the prompt, including premodifying nouns (*text messaging, internet chat rooms*). This leads to a response that is very reliant on repetition from the prompt, which may be awkward at times (e.g., *This situation many times has the result of* ***many parents are worried***…)

In contrast, Excerpt 6 shows no use of language from the prompt:

Excerpt 6 File A_P2_63 Dimension 3 Score = -7.58

Note: Generic bundles are in bold and *it* is italicized.

And *it's* only logical to assume that electronic tools should be **a part of** the teaching methods, on condition that the role of the teacher is taken into consideration. **He or she** is the one who should dominate the class and not the means of his work. Children are often impressionable and easily distracted, always in need of a firm hand to guide them.

In this excerpt, the writer uses generic bundles (*he or she, a part of*) and it (*it's only logical*), and a great deal of variety in vocabulary choices (no repetition of content words).

Dimension 4: Overt suggestions. Dimension 4 is characterized primarily by the linguistic features that loaded negatively on this dimension (certainty verbs, necessity modals, process and group nouns), all of which contribute to the function of making overt suggestions.

Table 7: Linguistic features loading on Dimension 4, Overt suggestions

Feature	Factor Loading
Positive loadings	
Indefinite articles	0.45
Human nouns	0.38
Topical adjectives	0.38
Negative loadings	
Certainty verbs	-0.54
Process nouns	-0.54
Group nouns	-0.36
Necessity modals	-0.33

This language can be seen in Excerpt 7:

Excerpt 7 File B_P2_78 Dimension 4 score = -11.28

Note: necessity modals are in bold, certainty verbs underlined, and process and group nouns in italics.

Last but not least, students **should** be allowed to use technological tools at *school* because it will help them with their future *career*. Nowadays, in order to get a good *job* one **must** be computer literate. In short, everyone **must <u>know</u>** how to use a personal computer.

This excerpt illustrates the use of language in responses that make overt suggestions about the topic in question. It also highlights the use of certain nouns (e.g.,

job, career, school) that are not found in the texts that show more of the positive features.

The linguistic features that loaded positively on Dimension 4 include language that avoids such overt suggestions, and focuses on a different group of nouns (human nouns) as well as topical adjectives and indefinite pronouns. Excerpt 8 illustrates these features.

Excerpt 8 File B_P3_94 Dimension 4 score = +7.43

Note: human nouns are bolded, indefinite articles underlined, and topical adjectives are in italics.

To begin with tourism is an important source of profit for many countries. Countries with underdeveloped economy but with beautiful natural environment, for instance Jamaica, are one of the **tourists** favorite destination. **Tourists** boost the *local* economy as they consume the *local* products and use the *local* services such as accommodation. Moreover tourism is a mean of advertisement for the country, as they usually persuade others to visit the country and consequently they support the country financially.

While the writer of Excerpt 8 is still arguing a position in relation to the topic, she does not use the language of overt suggestions to persuade the audience. Rather, the argument is presented more neutrally and factually. The excerpt also contains more nouns such as tourists and adjectives such as *local*. These features are more associated with Prompt 3, as will be discussed more below.

Dimension 5: Stance vs. referential discourse. Dimension 5 also has positive and negative features. Texts that are characterized by stance have greater use of stance bundles, likelihood verbs + *that* clauses, present tense verbs, cognition nouns, and stance nouns + *that* clauses. On the other hand, texts that are characterized by the use of referential discourse include a greater proportion of referential bundles and prepositional phrases.

Table 8: Linguistic features loading on Dimension 5, Stance vs. referential discourse

Feature	Factor Loading
Positive loadings	
Proportion stance bundles	0.68
Likelihood verb + *that* clause	0.48
Present tense	0.39
Cognition nouns	0.34
Stance noun + *that* clause	0.32
Negative loadings	
Proportion referential bundles	-0.52
Prepositions	-0.41

Excerpt 9 below illustrates the use of stance. Stance bundles are in bold, stance complement clauses (likelihood verbs and noun + *that* clauses) are italicized, present tense is underlined, and cognition nouns are in small caps.

Excerpt 9 File D_P2_46 Dimension 5 score = +13.89

Nowadays, students of all ages use in class computers, electronic dictionaries, calculators, etc. However, some schools are of *the* ***opinion that*** *these kinds of tools* ***should not be*** *used in classrooms. …* **It is *believed*** *that students who daily use for instance, calculators in class do not activate their brain* and hence they are made passive. On the other hand, it is *thought that* students who use educational technology are more informed about what happen around the world, they become more active in the classroom and hence they gain general KNOWLEDGE.

In this excerpt, the use of stance bundles (*the opinion that, should not be, it is believed*) highlight the writer's and others' stance directly. The use of stance noun + *that* clauses (e.g., *the opinion that these…*) and likelihood verbs + *that* clauses (*it is thought that students…*) also help the writer to express opinions.

Excerpt 10, on the other hand exemplifies referential discourse, which provides a great deal of local linkages and more in-depth discussion of relationships over time and distance.

Excerpt 10 File A_P3_45 Dimension 5 score = -9.90

Tourism might have existed as a notion **in the past**, but the concept of a tourism industry is certainly post 20th century. With the advent of new technologies the movement of people between states has grown significantly, which in turn led to tourism (as it was called) generating a lucrative income for the receiving states. … Nowadays, with tourism on the rise due to the end of major wars in Europe, nations aggressively try to attract **more and more** visitors. The reasons are obvious; these visitors generate **a lot of** wealth for the local communities, increase the host nation's reserve of foreign currency and, if satisfied, improve the nation's image abroad.

Excerpt 10, in contrast to Excerpt 9 above, provides more informationally oriented discourse that is much more bound to specific and detailed facts rather than overt expressions of opinion.

Correlations with ECPE scores

As shown in Table 9, factor scores on four of the five dimensions were significantly correlated with holistic ECPE scores: literate vs. oral discourse (D1), prompt dependence vs. lexical diversity (D3), overt suggestions (D4), and stance vs. referential discourse (D5). Pearson *r* correlation coefficients for these four dimensions ranged from .11 to .43. Among them, literate vs. oral discourse dimension (D1) had the strongest correlation with the ECPE score (r = .43, p < .01). The positive and strong correlation provides supportive construct-related validity evidence for the ECPE because essays produced by higher scoring writers tend to display clearer features of written discourse than those produced by lower scoring writers. The prompt dependence vs. lexical diversity dimension (D3) showed the second strongest correlation with ECPE scores (r = -.27, p < .01). This dimension is a lexical dimension: the positive side of this dimension is defined by its reliance on language from the prompt and the negative side by its lexical variety. Therefore, the negative correlation coefficient suggests that higher scoring writers tend to display a wider range of vocabulary use and develop content beyond the prompt.

A weaker correlation was found between the dimension of stance vs. referential expressions (D5) and ECPE score (r = -.18, p < .01). This dimension features the use of more explicit stance expressions (i.e., stance lexical bundles, stance-related nouns and verbs) on the one side and referential expression bundles (and prepositions) on the other. Therefore, the negative correlation indicates that higher scoring essays tend to display less explicit stance expressions and more referential expressions.

Table 9: Correlations among dimension scores and ECPE score

	D1	D2	D3	D4	D5	ECPE score
D1	-	-0.24**	-0.16**	0.12**	-0.20**	0.43**
D2		-	0.18**	-0.19**	0.09*	-0.01
D3			-	0.16**	-0.02	-0.27**
D4				-	-0.10*	0.11*
D5					-	-0.18**
ECPE score						-

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

The correlation between the overt suggestions dimension (D4) and score was even weaker (r = .11, p < .05), and thus we do not consider it a strong relationship. However, this was a rather interesting result. The negative side of this dimension features the use of certainty verbs (e.g., *conclude, prove, show*), necessity modal verbs (e.g., *should, must, have to*), and process nouns (e.g., *application, meeting, balance*), whereas the positive side of this dimension features indefinite articles, human nouns (e.g., *family, parent, teacher, president*), and topic adjectives (e.g., *political, international, national*). Certainty verbs and necessity modals are used to provide explicit arguments and suggestions, which, given their negative loadings, suggests that higher scoring essays tend to be less dependent upon explicit devices to convey arguments and suggestions. In contrast, the different semantic categories of nouns reflect the different contexts (from the prompts) where arguments and suggestions are made. The significant correlation for this dimension makes sense in that the dimension partly aligns with the stance vs. referential expressions dimension (D5), suggesting that higher scoring essays become more implicit in formal features of argumentation. However, the dimension is arguably more related to prompt as it conveys the contexts for the arguments and suggestions. As shown in subsequent factorial MANOVAs (see the following section), variation in factor scores of this dimension is more associated with features that represent a particular type of stance expression: overt suggestion. And these features are more frequent in essays for

Prompt 2, which asks examinees to write about their suggestions on the use of technology in the classroom.

Finally, the second dimension, topic-related content (D2), had a close to zero correlation with ECPE score. The non-significant correlation can be interpreted as discriminant validity evidence for the ECPE because D2 features concrete nouns (e.g., technical nouns, place nouns) and size and time adjectives, which are closely associated with the prompt topics. This dimension, though reflecting the topical or content differences between essays across prompts, is not a core criterion used to rate the ECPE essays.

Overall, the correlations among dimension scores and holistic essay score presented supportive validity evidence for the ECPE, in that higher scoring essays tend to demonstrate higher awareness of written discourse, a wider range of lexical knowledge, and more implicit stance expressions.

Factorial MANOVA of dimension scores

A two-way MANOVA of the dimension scores among the ECPE essays revealed significant multivariate main effects for score level (Wilks' $\lambda = .62$, F (20, 919.66) = 7.02, $p < .001$, $\eta^2 = .12$) and prompt (Wilks' $\lambda = .09$, F (10, 554) = 127.66, $p < .001$, $\eta^2 = .69$). The interaction effect between score level and prompt was not statistically significant (Wilks' $\lambda = .85$, F (40, 1210.21) = 1.19, $p = .20$). The non-significant interaction effect suggests that prompt differences did not affect score differences. That is, score differences followed the same statistical trend regardless of the prompt to which test takers responded. Score and prompt are two different constructs and the impact of one is not conditioned upon the other.

Given the significance of the overall test, the univariate main effects were examined. Using the adjusted alpha level ($\alpha = .01$), significant univariate main effects for score level were obtained for D1, (F (4, 281) = 22.09, $p < .001$, $\eta^2 = .24$), D3 (F (4, 281) = 13.40, $p < .001$, $\eta^2 = .16$), and D5 (F (4, 281) = 5.71, $p < .001$, $\eta^2 = .08$). In addition, significant univariate main effects for prompt were obtained for D2 (F (2, 281) = 408.40, $p < .001$, $\eta^2 = .74$), D3 (F (2, 281) = 19.14, $p < .001$, $\eta^2 = .11$), and D4 (F (2, 281) = 224.76, $p < .001$, $\eta^2 = .61$).

Although prompt effects were statistically significant on three of the five dimensions, after consideration of the distributions in the means plots (see Figure 1) and closer examination of the essays themselves in relation to the dimensions, we suggest that the five dimensions can be divided more clearly into two subgroups: (1) writing proficiency dimensions, i.e., dimensions that reflect score differences more than prompt differences; and (2) prompt dimensions, i.e., dimensions that reflect prompt differences more than score differences.

The writing proficiency dimensions include literate vs. oral discourse (D1), prompt dependence vs. lexical diversity (D3), and stance vs. referential discourse (D5). These dimensions involve lexico-grammatical features that mark written discourse, lexical sophistication, and implicitness of argumentation. These dimensions also had the strongest correlations with the holistic ECPE score awarded by human raters. Although D3 also showed significant differences across prompts, the means plots illustrate a greater rise across score levels than across prompts. The effect size for score level (η^2 =.16) was greater than for prompt ($\eta^2 = .11$). The means plots for D3 also shows the overall linear drop across score levels (particularly from score level D to score level E). Therefore, we suggest that this dimension is still more aligned with score differences. However, we also acknowledge that the significant prompt difference on D3 likely results from the fact that one of the prompts, Prompt 1, was more likely to contain prompt-match bundles. This prompt contained key words such as *text messaging* and *internet chat* that were picked up and repeated by the test takers.

In comparison, the prompt dimensions, which include topic-related content (D2) and overt suggestion (D4), reflect content differences among the three prompts. D2 features the use of concrete nouns, technical nouns and place nouns to denote the issues of technology, future, and globalization. These issues are aligned with the clear topic distinction among Prompt 1 (i.e., communication among young people at the digital age), Prompt 2 (i.e., technology in the classroom), and Prompt 3 (i.e., tourism and economy). Although D4 correlates with ECPE scores, it was a weak correlation (r = .11), and the combination of correlation analysis and MANOVA tests suggests that this dimension is more associated with prompt differences. Specifically, labeling D4 as a prompt dimension is supported by two pieces of evidence. First, this dimension contains negatively loading variables, certainty verbs and necessity modals, which tend to be used to express explicit stance and suggestions. An examination of the prompts suggests that while Prompts 1 and 3 elicit an opinion on a particular issue, Prompt 2 requires an explicit stance (either in agreement or disagreement) with the proposal of an educational policy. This prompt difference explains

the positive dimension scores for Prompts 1 and 3 but negative dimension scores for Prompt 2. Second, this dimension has positively loading features of human nouns (e.g., *family, parent, teacher*) and topic adjectives (e.g., *political, international, economic*). These two types of lexico-grammatical features are relevant to Prompts 1 and 3, respectively; and can therefore explain the positive dimension scores of these two prompts.

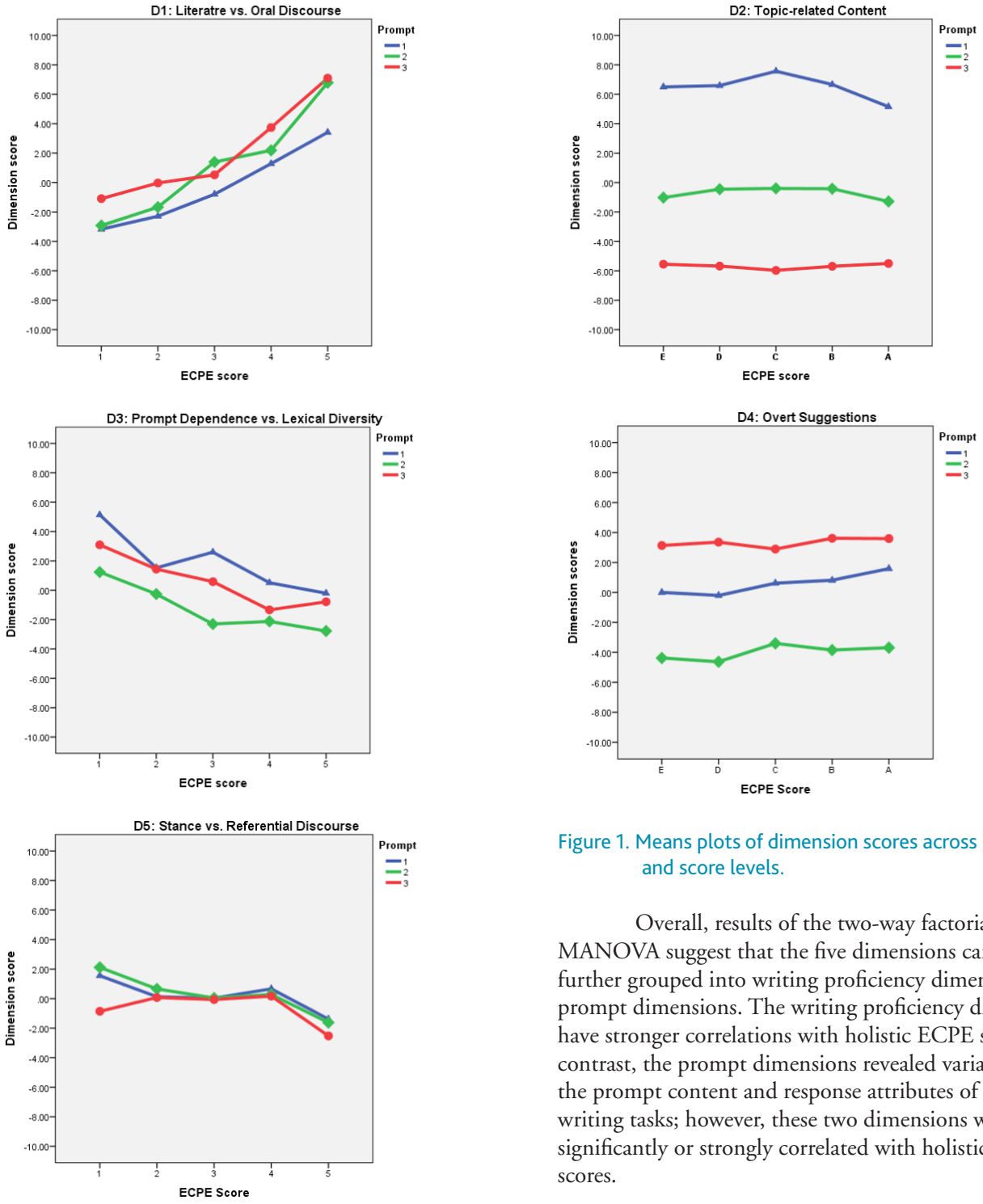

Figure 1. Means plots of dimension scores across prompt and score levels.

Overall, results of the two-way factorial MANOVA suggest that the five dimensions can be further grouped into writing proficiency dimensions and prompt dimensions. The writing proficiency dimensions have stronger correlations with holistic ECPE scores. In contrast, the prompt dimensions revealed variability in the prompt content and response attributes of the three writing tasks; however, these two dimensions were not significantly or strongly correlated with holistic ECPE scores.

Conclusion

This study investigated the lexico-grammatical features of ECPE writing performance through a MD approach. It also examined relationships between co-occurring linguistic features and the different levels of the ECPE writing scale. Specifically, we found five functional dimensions of lexico-grammatical features represented in the data set. Four of these functional dimensions (literate vs. oral discourse, prompt dependence vs. lexical variety, and stance vs. referential discourse, and overt suggestions) were significantly correlated with score level. However, among them, only three dimensions demonstrated significant differences across score levels (literate vs. oral discourse, prompt dependence vs. lexical variety, and stance vs. referential discourse). This suggests that these three dimensions were most representative of the construct of lexico-grammatical complexity as a core analytic component of writing proficiency measured by the ECPE writing tasks. In contrast, the other two dimensions (topic-related discourse and overt suggestions) were more closely related to prompt difference than score level as they demonstrated contrasts of content among the three prompts and their correlations with score level were either weak or non-significant.

The emergence of the three proficiency-related dimensions converges with previous literature. The literate vs. oral discourse dimension has been consistently observed as the strongest dimension in academic writing (e.g., Biber, 1988; Biber, 2006; Biber et al., 2016). Prompt dependence operationalized as a lexical dimension, or inability to develop content beyond what the prompt provides, was a marker of lower scoring essays in previous studies (e.g., Banerjee et al., 2015; Staples, Egbert, Biber & McClair, 2013). In addition, although they did not investigate proficiency, novice writers were found to use fewer referential bundles than expert writers in Chen and Baker (2010).

In addition, new to this study is the emergence of two prompt related dimensions. These two dimensions demonstrated significant differences, suggesting variation in ways that the prompt content elicits lexical choices (e.g., nouns, adjectives, verbs) among the essays to address the tasks. However, the fact that these prompt dimensions were less associated with score differences indicates that prompt differences did not lead to construct-irrelevant variances in the ECPE writing scores awarded by human raters. Taken together, both proficiency- and prompt-related dimensions provide supportive evidence for the construct-related validity for the ECPE, with particular respect to the scalability of lexico-grammatical complexity.

In a broad sense, the MD analysis demonstrated in this study provides a systematic approach to examining the scalability of lexico-grammatical complexity in writing assessment. Traditionally, lexico-grammatical complexity tends to be operationalized through either holistic measures (e.g., t-unit-based complexity measures) or single indicators of particular facets of lexico-grammatical complexity (e.g., noun of modifiers per noun). Given the multi-faceted nature of lexico-grammatical complexity, using only a few individual measures often leads to construct-underrepresentation and might not be sufficiently sensitive to differentiate between levels of writing proficiency (Banerjee et al., 2015; Biber et al., 2016). In contrast, the MD approach combines fine-grained and holistic analyses, looking at an array of individual linguistic features and reducing these features to a few linguistically functional dimensions to illustrate a fuller picture of the different layers in the use of lexical items and grammatical structures.

From the perspective of rater cognition, the MD approach may better reflect the rating process of lexico-grammatical features. That is, on a hybrid scoring rubric integrating analytic components in holistic scores, raters tend to examine the co-occurring patterns of multiple linguistic features rather than focusing on one or a few at a time. This makes "pure" analytic approaches to validation of holistic scales problematic in that raters do not typically (and should not be trained to) rate on single linguistic features. Therefore, fine-grained analysis of individual linguistic features alone may not reasonably reflect the rating process in writing assessments, especially when hybrid scoring rubrics are used.

Finally, this study linked observable lexico-grammatical features to abstract functional elements of academic writing, which bears important implications for rater training and rubric use. Most existing rubrics or rater training approach the scoring of functional features impressionistically. This approach is more likely to result in lower rater reliability on the scoring of functional features (e.g., argumentation, organization). Through a MD approach, the examination of individual lexico-grammatical features is conducive to a stronger link between formal features and functional elements in academic writing. In the case of this study, the three proficiency-related dimensions of lexico-grammatical features can be boiled down to three principles to guide raters in scoring: (1) responses should have more of

the characteristics of writing than speech; (2) responses should not contain too much of the prompt and should have lexical variety; (3) argumentative essays by higher proficiency writers become less marked by explicit rhetorical devices and more implicit in the formal features of argumentation. When representative formal features of different functions across score levels are exemplified in rater training materials, the training can help create alignment among raters in terms of the operationalization of linguistic functions.

References

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

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

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

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at …: Lexical bundles in university teaching and textbooks. *Applied Linguistics*, 25(3), 371-405.

Biber, D., & Gray, B. (2013). *Discourse Characteristics of Writing and Speaking Task Types on the TOEFL iBT Test: A Lexico-grammatical Analysis*. TOEFL iBT Research Report (TOEFL iBT-19). Educational Testing Service.

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

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

Cao, Y., & Xiao, R. (2013). A multi-dimensional contrastive study of English abstracts by native and non-native writers. *Corpora*, 8(2), 209-234.

Chen, Y. & Baker, P. (2010). Lexical bundles in L1 and L2 academic writing. *Language Learning and Technology*, 14(2), 30-49.

Davies, M. (2011). *Word and phrase.info*. Available online at www.wordandphrase.info.

Egbert, J. (2015). Publication type and discipline variation in published academic writing: Investigating statistical interaction in corpus data. *International Journal of Corpus Linguistics*, 20(1), 1-29.

Friginal, E., Li, M., & Weigle, S. (2014). Revisiting multiple profiles of learner compositions: A comparison of highly rated NS and NNS essays. *Journal of Second Language Writing*, 23, 1-16.

Jarvis, S., Grant, L., Bikowski, D., & Ferris, D. (2003). Exploring multiple profiles of highly rated learner composition. *Journal of Second Language Writing*, 12(4), 377-403.

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

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.

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

R Core Team (2015). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Simpson-Vlach, R., & Ellis, N. C. (2010). An academic formulas list: New methods in phraseology research. *Applied Linguistics*, 31(4), 487-512.

Staples, S., Egbert, J., Biber, D., & McLair, A. (2013). Formulaic sequences and academic writing development: Lexical bundles in the TOEFL iBT writing section. J*ournal of English for Academic Purposes*, 12(3), 214-225.

Staples, S., LaFlair, G., & Egbert, J. (2014). *Investigating the multi-faceted nature of speaking performance with a multivariate method*. Paper presented at the meeting of Midwest Association of Language Testers, Ann Arbor, MI.

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

Weigle, S. C. & Friginal, E. (2015). Linguistic dimensions of impromptu test essays compared with successful student disciplinary writing: Effects of language background, topic, and L2 proficiency. *Journal of English for Academic Purposes*, 18, 25-39.

Appendix

Coding Scheme for Generic Lexical Bundles

Functional categories
(adapted from Biber et al., 2004; Simpson-Vlach & Ellis, 2010)

1. Referential expressions
   a. Specification of attributes
      1. Intangible
      2. Tangible
      3. Quantity
   b. Identification of focus
   c. Contrast and comparison
   d. Deictics and locatives
   e. Vagueness markers
2. Stance expressions
   a. Epistemic stance
   b. Expression of ability and possibility
   c. Evaluation, obligation and directive
   d. Intention, volition and prediction
3. Discourse organizers
   a. Metadiscourse and textual reference
   b. Topic introduction and focus
   c. Topic elaboration
      1. Non-causal
      2. Cause and effect
   d. Discourse marker