Constructing Measurement Models of L2 Linguistic Complexity:  
A Structural Equation Modeling Approach

Takeshi KATO  
Graduate School, University of Tsukuba

Abstract
Over the last four decades, the constructs of complexity, accuracy, and fluency have been in focus in the analysis of language learners’ performance. However, due to the polysemous nature of complexity, more and more sub-constructs have been assumed, making holistic measurement difficult. This study aims to construct a more appropriate measurement model of L2 complexity by implementing finer-grained and relatively novel linguistic indices for capturing subordinate constructs that could not be measured by conventional indices. By utilizing five natural language processing tools, conventional and fine-grained indices of complexity were computed from 503 argumentative essays written by Japanese English learners. First, exploratory factor analysis was performed on linguistic index values and the extracted factor structures behind them. Second, confirmatory factor analysis was conducted to confirm whether the structure fits the data. Finally, a structural equation model of complexity constructs to predict essay scores was tested to evaluate its applicability to writing evaluation. The result of a series of factor analyses showed that the extracted factor structures reasonably fitted to the data for syntactic complexity (CFI = .901 and RMSEA = .071) and for lexical complexity (CFI = .978 and RMSEA = .051). Furthermore, the result of Structural Equation Modeling (SEM) analysis, which was proposed as a predictive model, accounted for 32.3 % of the variance of essay scores (CFI = .916 and RMSEA = .077). Overall, the findings showed the effectiveness of the proposed approach, which combined conventional linguistic features with fine-grained and relatively novel indices.

Keywords: syntactic complexity, lexical complexity, construct, writing assessment, automated essay scoring

Textual features have been considered and utilized as a link between language learners’ proficiency and their performances of target languages in language assessment and developmental language studies (Wolfe-Quintero, Inagaki, & Kim, 1998). Over the past 45 years, the complexity, accuracy, and fluency (CAF) triad has been the most frequently used framework in this research area (Housen, Kuiken, & Vedder, 2012). However, the polysemous nature of complexity has caused several methodological issues regarding the
measurement of the construct (Bulté & Housen, 2012; Pallotti, 2009). In both written and spoken language analyses, complexity has generally been divided into two aspects: syntactic complexity and lexical complexity (Sakuragi, 2011; Wolfe-Quintero et al., 1998). Problems in measuring these two constructs are outlined below.

As for syntactic complexity, large-grained linguistic indices represented by mean length of T-unit were most frequently used in its measurement (Ortega, 2003). However, these indices are difficult to interpret, since it is not clear what kind of complexification is operationalized in language performance (Ortega, 2012). In addition, Biber, Gray, and Poonpon (2011), who comprehensively investigated the linguistic features of a large-scale English speaking and writing corpora using Biber Tagger, reported that linguistic indices that discriminate proficiency in writing are phrase-level elaboration (e.g., attributive adjectives and prepositional phrases as modifiers for nouns). According to the conceptual model of syntactic complexity proposed by Bulté and Housen (2012, p. 27), this construct reflects syntactic sophistication and diversity. Wolfe-Quintero et al. (1998) also define this construct as “a wide and variety of both basic and sophisticated structures [that] are available to the learner” (p. 69). However, such norms (i.e., diversity and sophistication) have rarely been investigated (Yoon, 2017).

Regarding lexical complexity, learners’ lexical diversity and sophistication have been operationalized into linguistic indices for a single word, but this construct does not only consist of lexemic complexity; it also involves collocational complexity at the theoretical level (Bulté & Housen, 2012). The number of studies that focus on the frequency and range of multiword units has recently been increasing (e.g., Biber et al., 2004). In accordance with these research trends, newer indices that consider n-gram frequency and range have been developed and employed; the relationship between n-gram frequency and analytic scores of lexical proficiency (Kyle & Crossley, 2015) and correlations between n-gram range and human judgments of L1 independent writing tasks (Crossley, Cai, & McNamara, 2012) have been reported. In this way, lexemic complexity and collocational complexity have been recognized and measured individually, but they have not been empirically modeled as subordinate constructs of lexical complexity as is implied in Bulté and Housen’s (2012) definition.

To meet the above academic needs, the Tool for Automated Analysis of Syntactic Sophistication and Complexity (TAAASSC; Kyle, 2016) and the Tool for Automated Analysis of Lexical Sophistication (TAALES; Kyle & Crossley, 2015) were developed. These analyzers have recently been utilized in L2 writing studies, and it is reported that with their predictive models, which include fine-grained and novel indices, it is possible to predict proficiency with higher accuracy than models using only conventional indices (e.g., Kim, Crossley, & Kyle, 2018; Kyle, Crossley, & Berger, 2018).

This study aims to construct more appropriate measurement models of complexity by combining conventional subordinate constructs of syntactic and lexical complexity, and sub-constructs that have not been measured (i.e., syntactic diversity and sophistication) or
modeled (i.e., range norms and n-gram sophistication).

**Literature Review**

**L2 Complexity as a Multidimensional Construct**

Several previous studies have considered complexity as a conceptual construct. Ellis and Burkhuiizen (2005) defined it as “[the] use of more challenging and difficult language… Complexity is the extent to which learners produce elaborated language” (p. 139). Wolfe-Quintero et al. (1998) stated that “grammatical and lexical complexity mean that a wide and variety of both basic and sophisticated structures and words are available to the learner” (p. 69, p. 101), and Skehan (2003) defined complexity as “the complexity of the underlying interlanguage system developed” (p. 8). More recently, Housen et al. (2012) summarized the definitions offered by previous studies and proposed the following working definition: “the ability to use a wide and varied range of sophisticated structures and vocabulary in the L2” (p. 2). Researchers have attempted to capture these conceptual notions with linguistic features extracted from learners’ spoken and written performances.

According to Bulté and Housen (2012), complexity can take two forms: absolute complexity and relative complexity. The former refers to formal linguistic complexifications (e.g., the amount of subordination), while the latter indicates the relative difficulty of learning and processing particular linguistic structures related to input frequency and contingency.

Among the several subordinate constructs of complexity, linguistic complexity is the most-studied area of complexity in research exploring the developmental state of learners’ interlanguage by utilizing linguistic features in performance. This study focuses on syntactic and lexical complexity, both of which have been keenly adopted in numerous studies (e.g., Kato & Ono, 2019; Lu, 2010, 2012; Norris & Ortega, 2009; Wolfe-Quintero et al., 1998).

Figure 1 summarizes arguments made in previous researches and presents a structure of the subordinate constructs of linguistic complexity. The following sections describe syntactic complexity and lexical complexity in detail based on this figure.

![Figure 1](image-url)
Syntactic Complexity

Syntactic complexity, defined as “the range of forms that surface in language production and the degree of sophistication of such forms” (Ortega, 2003, p. 492), has been an important construct in studies on the development of L2 proficiency. Both systematic corpus studies (e.g., Biber et al., 2011; Lu, 2011) and comprehensive research syntheses (e.g., Norris & Ortega, 2009; Ortega, 2003; Wolfe-Quintero et al., 1998) have validated and verified the applicability of syntactic complexity measures to L2 writing measurement and development studies. Conventionally, it has been assumed that complex linguistic structures with varied types of embedded elements reflect learners’ proficiency levels based on language system development. Therefore, several global or clause-level indices of linguistic complexity have been widely employed in L2 writing research (Ortega, 2003). Clausal subordination measures have predicted L2 proficiency levels satisfactorily (Wolfe-Quintero et al., 1998) and clausal coordination has been used to predict proficiency of language development at beginning levels (Bardovi-Harlig, 1992).

Lexical Complexity

Many studies have investigated the vocabulary knowledge of L2 learners (e.g., Meara, 1996; Read, 2000). Most of those studies have assumed that the words in learners’ language usage reflect vocabulary knowledge and can be used to verify learners’ general L2 proficiencies (e.g., Daller & Xue, 2007; Jarvis, 2013; Lu, 2012; Saito, Webb, Trofimovich, & Isaacs, 2016).

Multiple technical terms have been devoted to productive vocabulary knowledge, the most common of which is lexical richness. The concept captures the overall quality of the vocabulary in a linguistic sample (Daller, Van Hout, & Treffers-Daller, 2007; Milton, 2009; Read, 2000). Many researchers have followed Read (2000) in using the concept of lexical richness as a superordinate construct to cover all lexical constructs and their associated measures.

In CAF-based research, lexical richness has been treated generally as lexical complexity since the target subordinate constructs and linguistic phenomena are shared. Bulté and Housen’s (2012, p. 28) model of lexical complexity includes lexical density, sophistication, and diversity, all of which were listed in Read’s (2000) sub-categories of lexical richness.

Expanding the Coverage of Complexity Measurement

As shown in Figure 1, complexity is a construct that implies syntactic and lexical diversity and sophistication. Previous studies have focused on structural complexity for syntactic complexity and word-level complexity for lexical complexity.

However, regarding syntactic complexity, the diversity and sophistication of syntactic structures have rarely been measured (Yoon, 2017). Moreover, large-grained indices such as length-based indices have been criticized regarding the difficulty of interpreting them (Norris...
For instance, Biber et al. (2011) reported that phrasal complexification is a more appropriate feature of academic writing than clausal complexification through comprehensive corpus analysis. Regarding lexical complexity, collocational complexity has rarely been assumed, but recent studies have used relatively novel indices that capture n-gram diversity and sophistication (e.g., Biber et al., 2004; Crossley et al., 2012; Kyle & Crossley, 2015).

Considering those concepts are originally supposed to be added to the measurement model of complexity, as part of the target construct it is worthwhile to construct and validate expanded measurement models of syntactic and lexical complexity. For these purposes, this study implements two state-of-the-art analyzers and extended sub-constructs explained below.

**Tool for Automated Analysis of Syntactic Sophistication and Complexity**

Kyle (2016) created TAASSC to analyze fine-grained syntactic features of learners’ performance, such as fine-grained noun phrase complexity, fine-grained clausal complexity, and syntactic sophistication and variation.

**Fine-grained noun phrase complexity.** TAASSC provides 66 types of noun phrasal indices and they can be divided into three types. The first type calculates the average number of dependents per each phrase type (e.g., nominal subjects) and for all phrase types. The second one quantifies the occurrence of particular dependent types (e.g., adjective modifiers) regardless of the type of noun phrase in which they occur. The final type computes the average occurrence of particular dependent types in particular types of noun phrases (e.g., adjective modifiers occurring in nominal subjects). Additionally, for the first type of indices, a standard deviation is calculated as a variation of phrasal structures.

**Fine-grained clausal complexity indices.** TAASSC has 31 fine-grained clausal complexity indices. Both finite and non-finite clauses are considered clauses by TAASSC. To prevent structures that inherently include more words (e.g., prepositional phrases) receiving more weight than those that do not (e.g., adjectives), TAASSC counts the length of clauses as the number of direct dependents per clause. Additionally, instead of grouping structures such as dependent clauses or complex nominals, TAASSC counts each type separately. Of the 31 indices, 29 calculate the average number of particular structures per clause while two calculate the total number of dependents per clause.

**Syntactic sophistication and variation.** In terms of cognitive linguistics, TAASSC operationalizes syntactic sophistication and variation with frequency scores and TTR of Verb Argument Construction (VAC) offered by the Corpus of Contemporary American English (COCA; Davies, 2009, 2010).

VAC in TAASSC is defined as a main verb and all the direct dependents it requires. In the sentence *Taro eats fast*, the main verb is *eats*, which takes two direct dependents: *Taro* and *fast*. *Taro* is the subject of the verb *eats*, and *fast* is an adverb that modifies *eats*. The clause *Taro eats fast* can be represented by the VAC nominal subject – verb – adverbial modifier
(Kyle, 2016). VAC frequency indices are calculated for main verb lemmas, VACs, and their combinations. Frequency indices comprise the average frequency score of the target structure (e.g., a VAC) in a particular text. If a structure that occurs in a text does not exist in the reference corpus, then it does not count toward the index score. These indices indicate the frequency of the linguistic structures in a text in reference to their use in COCA.

TTR is also computed for main verb lemmas, VACs, and their combinations. A token count includes the total number of occurrences of a particular structure in a text. A type count consists of the total number of unique instances of a particular structure. These indices measure the diversity of syntactic structures used in a text.

These three categories of fine-grained and novel linguistic features have been employed recently in studies to predict L2 writing quality with human judgment. Such research has reported that the models outperformed traditional predictive models in terms of the proportion of variance explained (e.g., Kyle & Crossley, 2017, 2018).

**Tool for Automated Analysis of Lexical Sophistication**

Kyle and Crossley (2015) developed TAALES, which provides indices of the compensatory notion of simple frequency (i.e., range) and multiword units (i.e., bigrams and trigrams). In this section, those categories, which are related to conventional lexical complexity constructs, are discussed.

**Range as a compensatory norm for frequency.** Because frequency measures are based on the number of times a word occurs in a corpus as a whole, they are not sensitive to how widely a particular word is used. Range norm indices operationalize how widely a word is used in an input text, usually by counting the number of documents in which the word occurs in a reference corpus. Range indices have been implemented in predictive models to estimate essay writing scores with human judgment and have shown a significant association with the scores (e.g., Kim et al., 2018; Kyle et al., 2018).

**N-gram frequencies.** Many researchers have begun to shift their focus from single lexical items to larger units (e.g., Biber et al., 2004). Recently, n-grams have proven their usefulness in detecting cross-linguistic difference (e.g., Kyle, Crossley, Dai, & McNamara, 2013) and predicting writing quality (Crossley et al., 2012). Crossley et al. (2012), for example, calculated a variety of frequency-based indices for bigrams and trigrams in written and spoken sub-corpora in the BNC. They found that indices that measured bigram frequency, bigram proportions, and bigram accuracy were predictive of human judgments of essay quality.

In TAALES, each index calculates the mean word or n-gram frequency and range scores. This is achieved by dividing the sum of the frequency and range scores for the words or n-grams in a text by the number of words or n-grams in that text that receive the scores.

**Research Questions**

This study aims to construct more appropriate measurement models of complexity by
combining conventional subordinate constructs of syntactic and lexical complexity and sub-constructs that have not been measured or modeled. The current study is guided by the following two research questions (RQs):

RQ1: What linguistic indices, including fine-grained and novel indices, form factor structures in relation to conventional complexity indices?
RQ2: To what extent do the subordinate constructs of syntactic and lexical complexity contribute to the prediction of a writing score?

**Method**

**Essay Data**

A total of 503 essays authored by Japanese university students were analyzed. The writers’ English proficiency was estimated to range from the Common European Framework of Reference for Languages (CEFR) level A2 to B1 based on their TOEFL ITP score distribution ($M = 477.84$, $SD = 40.32$; Tannenbaum & Baron, 2011). The essays were written under the time limit of 40 minutes, using the Criterion® editor without a thesaurus or grammatical check functions; the target was set at 200 words ($M = 209.57$, $SD = 92.90$). The essays were collected in three independent experiments and different topics were chosen from the Criterion topic list each time. The three tasks were consistent in that they were argumentative tasks and were all at the same “college level 1st year” level.

Essays were scored by the automated scoring engine e-rater® (Attali & Burstein, 2006) developed by the Educational Testing Service ($M = 1.95$, $SD = 0.95$). This system is trained with essay data evaluated by well-trained professional raters using rubrics, and it automatically gives essays 1 to 6 points as a holistic rating. The correlation between the score of the e-rater and the rating of the human evaluator is reported as .97 (Attali & Burstein, 2006, p. 22).

**Tools and Indices**

TAASSC and TAALES were used to extract fine-grained and novel indices related to syntactic and lexical complexity. The former offers fine-grained noun phrase complexity indices, clausal complexity indices, and VAC indices. The latter provides indices that tap into the compensatory notion of simple frequency (i.e., range) and multiword units (i.e., bigrams and trigrams). In this study, indices meeting the definition of conventional linguistic complexity constructs were selected out of these three categories and used for analysis.

**Fine-grained phrasal complexity indices.** TAASSC provides 132 kinds of noun phrase complexity indices. Half of these count pronouns as noun phrases, while the other half do not. Noun phrases in English can consist of pronouns that do not take direct dependents. In order to avoid the confusion caused by skewness resulting from including indices involving pronouns, this study employs 66 indices that involve nouns as fine-grained phrasal complexity indices.
Fine-grained clausal complexity indices. This analytic tool computes 31 types of clausal complexity indices, all of which were included.

VAC frequency and variation. VAC frequency indices comprise the average frequency score of the target structure (e.g., a VAC) in a particular text. If a structure that occurs in a text does not exist in the reference corpus (i.e., COCA), it is not counted as the index score. Regarding syntactic variation, TTR are also computed. In this research, indices for main verb lemmas were excluded because these measures overlap lexical diversity indices.

Word frequency and range. Word frequency and range indices in TAALES are computed based on several corpora. Since this study deals with written texts, sub-corpora consisting of only written texts in BNC, COCA, the Lorge’s corpus of popular magazine articles (Thorndike & Lorge, 1944), and the Brown corpus (Kučera & Francis, 1967) were employed. Each index calculates the mean frequency and range scores by dividing the sum of the frequency and range scores for the words in a text by the number of words in that text that receive a frequency and range score.

N-gram frequency and range. Frequency and range indices for bigrams and trigrams in TAALES are computed with BNC and COCA. Each index calculates the mean n-gram frequency and range scores by dividing the sum of the frequency and range scores for the n-grams in a text by the number of n-grams in that text that receive the scores. The proportion of n-grams in the target text that also occur frequently in the reference corpus is also computed.

Procedure

First, from the essay data, 183 index values (19 conventional syntactic complexity indices, 66 noun phrase complexity indices, 31 fine-grained clausal complexity indices, 24 VAC indices, seven conventional lexical complexity indices, eight word frequency and range indices, and 28 n-gram frequency and range indices) were calculated using L2 Syntactic Complexity Analyzer (SCA; Lu, 2010), Lexical Complexity Analyzer (LCA; Lu, 2012), Coh-Metrix 3.0 (McNamara, Graesser, McCarthy, & Cai, 2014), TAASSC, and TAALES.

Then, in order to avoid problems of multicollinearity, for combinations whose correlation coefficient between index values exceeded .90, only one index that has been used frequently in previous studies was sustained. Thereafter, the feasibility of factor analysis was tested and exploratory factor analysis was conducted. For the analysis, the psych package (Revelle, 2018) in the statistical development environment R (R Core Team, 2018) was utilized.

Confirmatory factor analyses (CFAs) were then conducted to determine whether the measurement models fit the data. As requisites for CFA, univariate and multivariate normality, parameter estimation method, model fit indices used, and sample size were confirmed. First, using the MVN package (Korkmaz, Goksuluk, & Zarasiz, 2014), univariate and multivariate normality were judged based on skewness and kurtosis of the variables and Mardia’s
normalized estimate and were violated. Thus, the diagonal weighted least square method was used for parameter estimation. One of the factor loadings from each factor was fixed to 1.00 for scale identification. The model fit was checked by CFI of .90 or above (Arbuckle & Wothke, 1995), RMSEA of .08 or below (Browne & Cudeck, 1993), and SRMR of .08 or below (Hu & Bentler, 1999). The dataset did not have missing values. The sample size exceeded 200, which is considered large according to Kline’s (2005) guidelines.

Finally, in order to verify the extent to which the measurement model of the subordinate constructs of linguistic complexity can predict L2 writing quality, Structural Equation Modeling (SEM) analysis was conducted. The premise of this analysis and the evaluation of goodness of fit were the same as CFA and the model was constructed by the subordinate constructs of L2 complexity. The distribution of Criterion scores was positively skewed, and the number of grades was small; thus, it was treated as ordinal scale data.

Results

Exploratory Factor Analysis

Before performing the analysis, observational variables that do not meet the criteria were excluded on the basis of the following principles: (i) rarely appear in the participants’ responses, or (ii) extremely similar to any one of the remaining variables. Based on the shapes of distribution of each index and descriptive statistics, 50 indices that rarely occur in the submitted essays were excluded from the analysis (e.g., verbal modifiers per indirect object). Referring to the Pearson product-moment correlation coefficients between the 133 index values, 26 pairs with a correlation coefficient exceeding .90 were confirmed. Therefore, 26 indices were excluded and a total of 107 indices (84 syntactic complexity indices and 23 lexical complexity indices) were employed for subsequent analyses.

Syntactic complexity. As for the feasibility of conducting factor analysis, KMO test and Bartlett’s sphericity test were performed. The KMO test score was .657 (“mediocre,” Kaiser, 1974) and was judged to be a valid value as a sample. In addition, the significance in Bartlett’s test ($\chi^2 = 6877$, $df = 276$, $p < .001$) indicated the correlation between variables rejected the null hypothesis. The number of factors to be extracted was set to three based on the VSS standard (Revell & Rocklin, 1979). As to factoring method, weighted least square method was used because almost all the variables were not normally distributed and multivariate normality was not confirmed based on Mardia’s multivariate kurtosis. Sixty items were removed from the model on the basis of the following predetermined criteria. Indices were removed from the model if they did not have primary factor loadings that were $\geq .40$ or if the items loaded on more than one factor. A summary of the analysis is presented in Table 1.

The first factor seemed to capture phrasal complexity. This factor included 11 indices that operationalize phrasal elaboration, nominal modification, and their variation. Thus, this factor was interpreted as Phrasal Complexity. The second factor seemed to capture VAC frequency. Four positively loaded indices operationalize the frequency of VAC or verb-VAC
combinations. Negatively loaded items were related to variation of VACs or verb-VAC combinations and grammatical structures per clause. A text that earns a high score for this factor will tend to include more frequent VACs or verb-VAC combinations, more repetitive use of those constructions, and less grammatical modifiers such as adverbs and prepositions. Thus, this factor was interpreted as VAC Sophistication. The third factor seemed to capture clausal complexification. All of the four indices loaded on this factor quantify clausal subordination and coordination; thus, this factor was interpreted as Clausal Complexity. To verify the internal consistency of each factor, Cronbach’s coefficient alpha (Cronbach, 1951) were computed: for Factor 1, α = .84; for Factor 2, α = .84; and for Factor 3, α = .83.

**Lexical complexity.** The KMO test score was .767 (“middling,” Kaiser, 1974) and was judged to be a valid value as a sample. In addition, the significance in Bartlett’s test ($\chi^2 = 7872.5$, $df = 120$, $p < .001$) indicated the correlation between variables rejected the null hypothesis. The number of factors to be extracted was set to three based on the VSS standard (Revell & Rocklin, 1979). As mentioned earlier, the distribution of the variables was unknown and multivariate normality was not confirmed, the weighted least square method was utilized for factoring. Seven items were removed from the model based on the following predetermined criteria. Indices were removed from the model if they did not have primary factor loadings that were ≥ .40 or if the items loaded on more than one factor. A summary of the analysis is presented in Table 2.

The first factor seemed to capture single word frequency. The rareness of word occurrences was evaluated with simple frequency and range scores derived from three different reference corpora. This factor was interpreted as Word Frequency. The second factor seemed to capture lexical diversity. All indices loaded on this factor reflect word variety computed with a whole sequential thread of words or random sampled words in a text. This factor was interpreted as Lexical Diversity. The third factor seemed to capture n-gram frequency. Of the four indices loaded on this factor, three indices quantify the frequency of bigram or trigram use in a text with reference corpora. This factor was interpreted as N-Gram Frequency. To verify the internal consistency of each factor, Cronbach’s coefficient alpha (Cronbach, 1951) were computed: for Factor 1, α = .93; for Factor 2, α = .93; and for Factor 3, α = .76.

**Confirmatory Factor Analysis**

In order to verify whether the above-mentioned syntactic complexity and lexical complexity factor structures fit to the data, CFA was performed using the lavaan package (Rosseel, 2012) in R (R Core Team, 2018). The diagonal weighted least square method was applied to estimate parameters. Inter-variable Pearson product-moment correlations ($r = -.60$ to $.79$) for syntactic complexity and ($r = -.67$ to .89) were not high enough to cause multicollinearity ($r = .90$ or above; Tabachnick & Fidell, 2014).
Table 1  
Summary of Factor Loadings for Oblique Geomin-Rotated Three-Factor Solution With Conventional and Fine-Grained Syntactic Complexity Indices

<table>
<thead>
<tr>
<th>Index</th>
<th>Code</th>
<th>Factor loading</th>
<th>Communalty</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>dependents per nominal</td>
<td>Ph1</td>
<td>.88</td>
<td>-.04</td>
</tr>
<tr>
<td>CN/C</td>
<td>Ph8</td>
<td>.65</td>
<td>.01</td>
</tr>
<tr>
<td>dependents per nominal SD</td>
<td>Ph5</td>
<td>.62</td>
<td>-.07</td>
</tr>
<tr>
<td>MLC</td>
<td>Ph9</td>
<td>.61</td>
<td>.13</td>
</tr>
<tr>
<td>prepositions per nominal</td>
<td>Ph11</td>
<td>.58</td>
<td>-.03</td>
</tr>
<tr>
<td>dependents per direct object</td>
<td>Ph2</td>
<td>.54</td>
<td>-.04</td>
</tr>
<tr>
<td>dependents per nominal subject</td>
<td>Ph4</td>
<td>.53</td>
<td>.04</td>
</tr>
<tr>
<td>dependents per object of the preposition</td>
<td>Ph3</td>
<td>.52</td>
<td>.09</td>
</tr>
<tr>
<td>dependents per nominal subject SD</td>
<td>Ph7</td>
<td>.46</td>
<td>.01</td>
</tr>
<tr>
<td>dependents per object of the preposition SD</td>
<td>Ph6</td>
<td>.44</td>
<td>.05</td>
</tr>
<tr>
<td>adjectival modifiers per nominal</td>
<td>Ph10</td>
<td>.44</td>
<td>.05</td>
</tr>
<tr>
<td>logged average VAC frequency</td>
<td>VAC1</td>
<td>.03</td>
<td>.84</td>
</tr>
<tr>
<td>dependents per clause</td>
<td>VAC7</td>
<td>.05</td>
<td>-.77</td>
</tr>
<tr>
<td>VAC type-token ratio - all</td>
<td>VAC3</td>
<td>-.04</td>
<td>-.71</td>
</tr>
<tr>
<td>logged average verb-VAC combination frequency</td>
<td>VAC4</td>
<td>-.02</td>
<td>.60</td>
</tr>
<tr>
<td>verb-VAC combination type-token ratio</td>
<td>VAC5</td>
<td>.05</td>
<td>-.57</td>
</tr>
<tr>
<td>% of verb-VAC combinations in text are in COCA</td>
<td>VAC6</td>
<td>-.09</td>
<td>.52</td>
</tr>
<tr>
<td>prepositions per clause</td>
<td>VAC8</td>
<td>.14</td>
<td>-.50</td>
</tr>
<tr>
<td>adverbal modifiers per clause</td>
<td>VAC9</td>
<td>-.03</td>
<td>-.50</td>
</tr>
<tr>
<td>average VAC frequency SD</td>
<td>VAC2</td>
<td>.15</td>
<td>.45</td>
</tr>
<tr>
<td>C/S</td>
<td>Cl1</td>
<td>-.03</td>
<td>.04</td>
</tr>
<tr>
<td>DC/C</td>
<td>Cl2</td>
<td>-.02</td>
<td>.03</td>
</tr>
<tr>
<td>VP/T</td>
<td>Cl3</td>
<td>.06</td>
<td>-.13</td>
</tr>
<tr>
<td>subordinating conjunctions per clause</td>
<td>Cl4</td>
<td>.03</td>
<td>.14</td>
</tr>
</tbody>
</table>

Factor correlations

|                                                    |        |
| Factor 2                                          | .12    |
| Factor 3                                          | -.01   |

Eigenvalue

|                                                    |        |
| % of variance                                      | 16%    |
| Cumulative %                                       | 16%    |

Note. N = 503; Factor loadings ≥ .40 are in boldface. AW = all words; CN/C = complex nominals per clause; MLC = mean length of clause; C/S = clauses per sentence; DC/C = dependent clauses per clause; VP/T = verb phrases per T-unit.
<table>
<thead>
<tr>
<th>Index</th>
<th>Code</th>
<th>Factor loading</th>
<th>Communalit</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>BNC Written Range AW</td>
<td>Wf2</td>
<td>.95</td>
<td>.04</td>
<td>-.02</td>
<td>.88</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BNC Written Frequency AW Logarithm</td>
<td>Wf1</td>
<td>.82</td>
<td>-.06</td>
<td>.25</td>
<td>.88</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LS1</td>
<td>Wf7</td>
<td>-.82</td>
<td>.04</td>
<td>.32</td>
<td>.67</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>COCA Academic Range AW</td>
<td>Wf5</td>
<td>.81</td>
<td>.05</td>
<td>.26</td>
<td>.80</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BNC Written Bigram Proportion 50k</td>
<td>Wf3</td>
<td>.79</td>
<td>.02</td>
<td>.16</td>
<td>.70</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kučera-Francis Register Range AW</td>
<td>Wf6</td>
<td>.75</td>
<td>.11</td>
<td>-.13</td>
<td>.49</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>COCA Academic Frequency AW Logarithm</td>
<td>Wf4</td>
<td>.72</td>
<td>-.05</td>
<td>.32</td>
<td>.77</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LS2</td>
<td>Wf8</td>
<td>-.62</td>
<td>.07</td>
<td>.25</td>
<td>.41</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HD-D</td>
<td>Div4</td>
<td>.00</td>
<td>.94</td>
<td>-.04</td>
<td>.90</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MTLD</td>
<td>Div2</td>
<td>-.03</td>
<td>.93</td>
<td>.05</td>
<td>.87</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MSTTR</td>
<td>Div3</td>
<td>.07</td>
<td>.89</td>
<td>-.02</td>
<td>.75</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UBER</td>
<td>Div1</td>
<td>-.13</td>
<td>.73</td>
<td>.01</td>
<td>.61</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>COCA Academic Bigram Range Logarithm</td>
<td>Ngf1</td>
<td>.26</td>
<td>.01</td>
<td>.77</td>
<td>.76</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BNC Written Bigram Frequency Logarithm</td>
<td>Ngf2</td>
<td>-.01</td>
<td>-.06</td>
<td>.76</td>
<td>.59</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>COCA Academic Frequency</td>
<td>Ngf4</td>
<td>.02</td>
<td>-.03</td>
<td>.57</td>
<td>.33</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>COCA Academic Trigram Range Logarithm</td>
<td>Ngf3</td>
<td>-.07</td>
<td>.01</td>
<td>.46</td>
<td>.20</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Factor correlations</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Factor 1</td>
<td></td>
<td>-.34</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Factor 2</td>
<td></td>
<td></td>
<td>-.12</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Factor 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Eigenvalue</td>
<td>5.22</td>
<td>3.12</td>
<td>2.26</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>% of variance</td>
<td>33%</td>
<td>20%</td>
<td>14%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cumulative %</td>
<td>33%</td>
<td>52%</td>
<td>66%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. N = 503; Factor loadings ≥ .40 are in boldface. AW = all words; LS1 = ratio of sophisticated word tokens; LS2 = ratio of sophisticated word types; HD-D = hypergeometric distributed D measure; MTLD = the measure of textual lexical diversity; MSTTR = mean segmental TTR; UBER = uber index.

**Syntactic complexity.** Firstly, CFA was performed for each factor obtained by EFA. Regarding Phrasal Complexity, correlations between uniqueness of indices operationalizing similar grammatical phenomena were estimated beforehand. Ph1, Ph2, Ph3, and Ph4 calculate the arithmetic means of the number of dependents within specific nominal structures. Since Ph1 subsumes Ph2, Ph3, and Ph4 in terms of the category of the target structure, correlations between Ph1 and Ph2, Ph3, and Ph4 were naturally assumed. Ph5, Ph6, and Ph7 calculate the standard deviation of the number of dependents within specific nominal structures. The same reason is true of correlations between these indices; thus, correlations between Ph5 and Ph6 and Ph7 were also assumed. Because Ph1, Ph3, and Ph4; and Ph5, Ph6, and Ph7 share the target linguistic structures, respectively, correlations between each pair were assumed. The
model showed reasonable fit to the data, where CFI = .959, RMSEA = .075, and SRMR = .072.

Regarding Clausal Complexity, a uni-factor model that was obtained from EFA was verified. The model showed great fit to the data, where CFI = .997, RMSEA = .044, and SRMR = .032.

Regarding VAC Sophistication, error covariances of indices operationalizing similar grammatical phenomena were estimated beforehand. VAC1 and VAC2 share the target structures (i.e., the number of VACs), and an error covariance between them was assumed. The pairs VAC1 and VAC4; and VAC3 and VAC5 quantify similar structures by the same calculation method and thus, error covariances between them were assumed. One of the error covariances between VAC1 and VAC4 was not significant ($p = .181$), so this path was removed. The final model showed a reasonable fit to the data, where CFI = .963, RMSEA = .077, and SRMR = .078.

Each of the three factors was verified based on the above uni-factor CFA and the three-factor solution tested. The initial model that simply consisted of three uni-factor structures did not fit the data. Based on modification indices, since the amount of subordination (e.g., relative pronouns) would influence the number of dependents per nominal, paths were drawn from Clausal Complexity to Ph5 and Ph7. In addition, since the amount of clausal subordination and coordination would affect syntactic variation, paths were added from Clausal Complexity to VAC3 and VAC5. The final model showed a reasonable fit to the data, where CFI = .901, RMSEA = .071, and SRMR = .079. As competing models, the following two models were tested: (a) a single factor model in which all the three latent variables are fully correlated; and (b) an uncorrelated three-factor model in which all latent variables are not correlated. As a result, the former estimation did not converge, and thus comparison was impossible. For the latter, the goodness of fit was as follows: CFI = .869, RMSEA = .081, and SRMR = .088. Therefore, the proposed correlated three-factor was the best. Figure 2 shows the results.
Figure 2. Final CFA model of supplemented syntactic complexity (N = 503). All of the path coefficients are standardized and significant (p < .001). Phrasal = Phrasal Complexity; Clausal = Clausal Complexity; VACsop = VAC Sophistication.

Lexical complexity. For Word Frequency, an error covariance between LS1 and that of LS2 was estimated because the calculation methods are quite similar. The model showed an excellent fit to the data, where CFI = .998, RMSEA = .027, and SRMR = .053. For N-Gram Frequency, the model showed great fit to the data, where CFI = .999, RMSEA = .000, and SRMR = .018. For Lexical Diversity, the model showed reasonable fit to the data, where CFI = .999, RMSEA = .000, and SRMR = .010.

Figure 3. Final CFA model of supplemented lexical complexity (N = 503). All of the path coefficients are standardized and significant (p < .001). Wfreq = Word Frequency; NGfreq = N-Gram Frequency; Diversity = Lexical Diversity.

Each of the three factors was verified based on the above uni-factor CFA and the three-factor solution was tested. For the three-factor solution of Lexical Complexity measurement model, the initial model that simply consisted of three uni-factor structures did not fit the data. Based on modification indices, since n-gram frequencies affect word
frequencies, paths from N-Gram Frequency to Wf7 and Wf8 were set. The final model showed a reasonable fit to the data, where $\text{CFI} = .978$, $\text{RMSEA} = .051$, and $\text{SRMR} = .072$. To compare the goodness of fit of potential models, as in the syntactic complexity, a single factor model and an uncorrelated three-factor model were tested. As a result, again, the former did not converge. Regarding the latter, the goodness of fit was as follows: $\text{CFI} = .718$, $\text{RMSEA} = .181$, and $\text{SRMR} = .191$. Thus, the correlated three-factor was the most promising model. Figure 3 summarizes the results.

**Structural Equation Modeling**

Finally, a model that predicts Criterion scores with subordinate constructs of syntactic and lexical complexity obtained from the previous analyses was built and tested. First, six latent variables were utilized as independent variables for predicting Criterion scores. However, four paths from VAC Sophistication ($B = .078, p = .069$), Word Frequency ($B = -.073, p = .108$), N-Gram Frequency ($B = .065, p = .166$), and Lexical Diversity ($B = .073, p = .175$) to Criterion scores were not significant, and these paths were removed. All paths in the final model then became significant. The model showed a reasonable fit to the data, where $\text{CFI} = .916$, $\text{RMSEA} = .077$, and $\text{SRMR} = .079$. Figure 4 shows the results.

Two regression paths from Clausal Complexity ($B = .476, p < .001$) and Phrasal Complexity ($B = .336, p < .001$) were significant and the latent variables positively associated with Criterion scores. The predictive SEM model accounted for 32.3% of the variance of Criterion scores.

**Discussion**

**Factor Structures Behind Conventional and Fine-Grained Indices**

The first purpose of this study is to investigate the factor structures of extended syntactic and lexical complexity measured by conventional and fine-grained linguistic features. Results of EFA and CFA indicate that both constructs would consist of three subordinate constructs.
Figure 4. Structural equation model with supplemented syntactic and lexical complexity sub-constructs as predictors of Criterion scores ($N = 503$). All of the path coefficients are standardized and significant ($p < .001$). CS = Criterion scores.

As for lexical complexity, Word Frequency, N-Gram Frequency, and Lexical Diversity were extracted as common factors. The Word Frequency factor includes conventional lexical sophistication indices (i.e., Wf7: LS1 and Wf8: LS2), and this factor would thus be in line with the norm of lexical sophistication. Inter-factor correlation between this construct and N-Gram Frequency were moderate ($r = .50$); the N-Gram Frequency factor extends to the traditional lexical sophistication norm. This result supports Kim et al. (2018), who distinctly extracted principle components of “Content Word Frequency” and “Bigram Frequency and Range.”

Writing Score Prediction With Supplemented L2 Complexity Sub-Constructs

The second aim of this study is to explore the degree of predictive performance for writing quality of latent variables related to linguistic complexity. The results of SEM suggest that supplemented Phrasal Complexity and Clausal Complexity are significantly associated with Criterion scores and accounted for 32.3% of variance of the score. The proportion of variance explained is larger than that of the model constructed by only conventional sub-constructs of linguistic complexity (29.8%; Kato & Ono, 2019). This result supports Biber et al.’s (2011) proposal regarding the necessity of considering phrase level linguistic elaborations, specifically noun phrase elaborations. In addition, the trend that the combined
(i.e., both conventional and fine-grained indices used) predictive model accounts for a larger portion of variance of the writing score than a model consisting only of conventional indices has been reported by Kyle (2016) and Kyle and Crossley (2017).

On the other hand, four out of six latent variables are not associated with writing quality derived from an automated essay scoring system. Regarding Lexical Diversity, too strong robustness against length would lead to less association with essay scores. The other three latent variables mainly influence the use of common specific linguistic items (i.e., VAC Sophistication, Word Frequency, and N-Gram Frequency), which are opposite concepts from the previous approaches that respect low frequency. Although several studies confirm significant relationships between writing quality and VAC frequency (Kyle, 2016; Kyle & Crossley, 2017), and Word Frequency and N-Gram Frequency (Crossley et al., 2012; Kim et al., 2018; Kyle et al., 2018), the results of this study are different. There is a possibility that the two analyzers compute indices that tap into those constructs with reference to native speakers’ corpus (e.g., BNC and COCA), the proficiency level of writers or participants in the previous studies were higher than those in this study, and thus there is greater variance in the frequency scores. This may have contributed to the predictions in regression analyses.

**Conclusion and Future Research**

In this study, in order to measure larger coverage of vast complexity constructs, attempts were made to integrate conventional indices and subordinate constructs of syntactic complexity and fine-grained and novel linguistic indices. Extracted factors were interpreted in line with previously proposed complexity constructs, and a predictive model consisting of these factors outperformed the model proposed in previous studies. The results also showed the direction of validating the measurement model of complexity (i.e., introducing finer-grained indices into the measurement model regarding theoretical backing) and indicated the necessity to develop more reliable indices for capturing EFL writers’ complexification in their performance.

As with most research, this study has a number of limitations. First, all of the authors of essays used as data in this study belong to a single educational institution, which may limit the generalizability of the results. Additionally, their writing proficiency scores were relatively low and may cause a skewed distribution of the Criterion scores. While this study employed statistical methods that have robustness against skewed data, it may be necessary to include essays written by more proficient language learners in order to obtain more valid results.

Another limitation that could be addressed in future studies is the source of the writing scores. In this study, the e-rater was employed as the sole rater to assign scores to each essay, but this tool does not fully reflect human judgment. Thus, a more detailed exploration of the relationship between human perception and conventional and fine-grained complexity indices on essays would be necessary.
Acknowledgements

I would like to thank Professor Yuichi ONO and my colleagues for their helpful advice. I extend my gratitude to two anonymous reviewers for their insightful comments to improve this paper.

References


Davies, M. (2010). The Corpus of Contemporary American English as the first reliable


