Continuing to Explore the Multidimensional Nature of Lexical Sophistication: The Case of Oral Proficiency Interviews

MASAKI EGUCHI[1,2] and KRISTOPHER KYLE[2,3]
[1] *University of Hawai'i at Mānoa, Department of Second Language Studies, Moore Hall, 1890 East West Road, Honolulu, HI, 96822 Email: masakie@uoregon.edu*
[2] *University of Oregon, Department of Linguistics, 161 Straub Hall, 1290 University of Oregon, Eugene, OR, 97403-1290 Email: kkyle2@uoregon.edu*
[3] *Yonsei University, English Department, 50 Yonsei-ro, Sinchon-dong, Seoul, 03722, South Korea*

Lexical sophistication has been an important indicator of productive lexical proficiency for almost 30 years. Although lexical sophistication has most often been operationalized as the proportion of low frequency words in a text, a growing body of research has indicated that a number of indices such as concreteness, hypernymy, and n-gram association strengths meaningfully contribute to the construct. While the increase in available indices has expanded our understanding of the multidimensional construct, the sheer number of indices presents a practical barrier for researchers. Although some studies have begun to address this issue, most have been confined to the analysis of argumentative tasks, which are not necessarily representative of the range of tasks learners may encounter. This study therefore investigates the structure of lexical sophistication indices in a large learner corpus of English second language (L2) oral proficiency interviews (OPIs). An exploratory factor analysis identified 10 factors, 7 of which explained approximately 58% of the variance in OPI scores in a follow-up regression analysis. The results suggest that while some features of lexical sophistication (e.g., concreteness) may be task independent, others (e.g., frequency) may be task specific.

Keywords: lexical richness; oral proficiency interviews; exploratory factor analysis; corpus linguistics; natural language processing; vocabulary

THE PAST TWO DECADES OF SECOND language (L2) learning research have witnessed increasing attention to the measurement of learners' vocabulary knowledge and use (Daller, Milton, & Treffers–Daller, 2007; Read, 2000). Vocabulary is a critical dimension of language use both in theoretical psycholinguistic models of L2 production (Kormos, 2006; Skehan, 2009) and in empirical investigations of L2 performance (de Jong et al., 2012; Koizumi & In'nami, 2013). Researchers generally agree that the quality of lexical resources will determine the extent to which learners achieve their functional goals in communicative contexts (Crossley, Salsbury, & McNamara, 2015; Lu, 2012; Saito et al., 2016b).

In an attempt to characterize vocabulary use, the notion of lexical richness has been defined and operationalized using a set of subordinate concepts—primarily, lexical diversity and lexical sophistication—both of which in turn comprise a number of subdimensions (Bulté & Housen, 2012; Crossley et al., 2011b; Lu, 2012; Nation & Webb, 2011; Read, 2000; Skehan, 2009). Lexical diversity, as the name suggests, concerns the variety of words demonstrated in a text and has seven subdimensions (see Jarvis, 2013). Among them, lexical variability is currently by far the most commonly operationalized construct of diversity.

The Modern Language Journal, 104, 2, (2020)
DOI: 10.1111/modl.12637
0026-7902/20/381–400 $1.50/0

Lexical variability refers to the variety of unique word forms used in a text, and is often operationalized using a variant of the type–token ratio (McCarthy & Jarvis, 2010; Treffers–Daller, Parslow, & Williams, 2016). The working hypothesis of lexical diversity is that higher proficiency language users will use a wider variety of vocabulary items. Lexical diversity can reflect the size of a lexicon (i.e., how many words individuals know; Nation, 2013), but most of its subconstructs do not concern the characteristics of those vocabulary items (Bulté & Housen, 2012).[1]

Complementary to this notion is lexical sophistication, the focus of the current article, which attempts to characterize "the proportion of relatively unusual or advanced words in the learner's text" (Read, 2000, p. 203; see also Laufer & Nation, 1995). Although word frequency has often been used to operationalize 'unusual' or 'advanced' lexis, multiple indices beyond frequency can be used to measure sophistication (e.g., hypernymy, concreteness, multiword units; Crossley, Salsbury, & McNamara, 2009; Kim, Crossley, & Kyle, 2018; Kyle, Crossley, & Berger, 2017), depending on the researcher's definition (see Laufer & Nation, 1995; Meara & Bell, 2001). Lexical sophistication can not only tap into vocabulary size (i.e., learners with larger mental lexicons will produce more sophisticated words), but also aspects of vocabulary depth, particularly the use dimension of Nation's (2013) vocabulary knowledge framework (i.e., grammatical functions and collocations; Bestgen & Granger, 2014; Kyle et al., 2017; Schmitt, 2014).

Despite the centrality of frequency-based sophistication in L2 research (Bulté & Housen, 2012; Michel, 2017), recent studies have demonstrated that a number of lexical sophistication indices can be used to measure sophistication from multiple angles (Kyle et al., 2017). This is a desirable movement because while useful, frequency measures tend to underrepresent the multifaceted nature of vocabulary knowledge and use (Daller et al., 2007; Koizumi, 2012; Nation, 2013). With the recent proliferation of lexical indices, however, it is becoming even more difficult for L2 researchers to understand the degree to which these indices measure distinct areas of lexical performances (Bulté & Housen, 2012; Norris & Ortega, 2009). To address this gap, Kim et al. (2018) took a preliminary step in this direction by reporting 12 lexical sophistication dimensions in a large corpus of argumentative L2 writings. The result was largely promising as it showed interrelated subdimensions of the construct. However, as previous studies have demonstrated potential task effects on lexical sophistication (Kyle & Crossley, 2016; Kyle, Crossley, & McNamara, 2016), it is unclear whether these dimensions extend to other contexts. In particular, lexical sophistication research has been primarily focused on argumentative written tasks, and has tended not to investigate oral proficiency interviews (OPIs), which are commonly used in educational and assessment settings. Therefore, this study conducted an exploratory factor analysis (EFA) followed by a multiple regression to examine the degree to which components of lexical sophistication are predictive of OPI scores.

FREQUENCY AS AN INDICATOR OF LEXICAL SOPHISTICATION

Traditionally, lexical sophistication has been operationalized with regard to the reference-corpus frequency of lexical items in a text. Well-known measures include the lexical frequency profile (Laufer & Nation, 1995), and mean frequency scores such as those measured by Coh-Metrix (Graesser et al., 2004) and the Tool for the Automatic Analysis of Lexical Sophistication (TAALES; Kyle et al., 2017). An underlying assumption of this approach is that acquisition will occur in order of frequency, suggesting that a higher proportion of lower frequency words in a learner text is a hallmark of a more elaborated mental lexicon (Bulté & Housen, 2012; Michel, 2017; Nation & Webb, 2011).

Despite the extensive applications of frequency for the past two decades, researchers have been aware of the need for alternative approaches for measuring advanced lexical use (Crossley et al., 2011b; Meara & Bell, 2001). Consequently, there is a growing consensus that the construct of lexical sophistication should be operationalized from different perspectives (Crossley et al., 2011a; Kyle & Crossley, 2015). Simply put, this research has suggested that lexical sophistication is a multidimensional phenomenon that should be measured as such (see Kim et al., 2018).

MULTIDIMENSIONAL LEXICAL SOPHISTICATION

In response to the growing need for a multidimensional operationalization, we conceptualize lexical sophistication under two related notions: relative complexity (i.e., the relative difficulty of learning, and therefore using, linguistic items; Bulté & Housen, 2012) and intralexical learning burden (i.e., the inherent difficulty of a lexical item due to its features such as frequency,

imageability, polysemy, and register restriction; Ellis & Beaton, 1993; Laufer, 1997). These two concepts (one from linguistic complexity research and the other from vocabulary research) collectively provide the theoretical bases of lexical sophistication. Multiple lexical features such as frequency, imageability, and formal complexity (among others) contribute to the difficulty of learning (and therefore using) a particular word or word combination.[2] Framing lexical sophistication this way, recent studies have demonstrated that various lexical features originally developed in a number of related disciplines (e.g., corpus linguistics, educational psychology, psycholinguistics) can be used to capture different aspects of sophisticated lexical use (Kim et al., 2018; Kyle et al., 2017). These emerging categories of indices are reviewed.

Word Range

One of the limitations of the frequency index is that specialized words repeated in a small number of documents can skew the frequency distribution in the reference corpus; therefore, it may not reflect how often a learner encounters a word in general language use. Range can overcome this by counting the number of documents in which a word occurs (Kyle & Crossley, 2015). Words that are used in a wide range of documents in the reference corpus are considered to be common or general words, while words that occur in fewer documents may be more register-specific, specialized vocabulary. Range has been shown to be an important predictor of holistic scores of L2 speaking proficiency, L2 written lexical proficiency (Kyle & Crossley, 2015), and writing quality (Kyle & Crossley, 2016).

Psycholinguistic Word Information

Lexical sophistication not only concerns the relative occurrence of lexical items in natural language (i.e., frequency and range), but also includes the conceptual image a word evokes (Salsbury, Crossley, & McNamara, 2011). Four related indices that are derived from subjective ratings by first-language (L1) English speakers (Coltheart, 1981) include concreteness (how concrete or abstract a word is), imageability (how easy it is to construct a mental image of a word), meaningfulness (how associated a word is to other words), and familiarity (how commonly a word is experienced). Indices based on these constructs have been found to reflect longitudinal development (Crossley & Skalicky, 2017; Salsbury et al., 2011), and have contributed to the variance in cross-sectional spoken performances (Crossley et al., 2011a; Saito et al., 2016a).

Age of Acquisition and Exposure

Age of acquisition (AoA) and age of exposure (AoE) refer to the average age at which a word is learned (AoA) or experienced (AoE) by L1 speakers (Kyle et al., 2017). These indices correlated with lexical decision latencies, suggesting that the indices reflect the processing difficulty of those lexical items (Kuperman, Stadthagen–Gonzalez, & Brysbaert, 2012).

Semantic Networks

It is a widely held view that proficient learners have a densely organized mental lexicon (Meara, 2005; Schmitt, 2014). Indices of semantic networks attempt to tap into this concept by measuring the use of polysemous and hypernymic words (Crossley et al., 2009; Crossley, Salsbury, & McNamara, 2010). The use of polysemous words has been indicative of the development of intricate meaning–form mappings (Crossley et al. 2010; Saito et al., 2016b). Hypernymy is considered to reflect the dense hierarchical relationships between words in the learner's lexicon (Crossley et al., 2009; Saito et al., 2016a). Both polysemy and hypernymy indices have been shown to be correlated with speaking and writing proficiency (Crossley et al., 2011a; Kim et al., 2018; Kyle et al., 2017).

Word Recognition Norms

Sophisticated vocabulary can also be operationalized in terms of the cognitive cost needed to access them. Word recognition norms utilize mean response latencies and accuracy scores of lexical decision and word naming tasks by L1 English speakers (Berger, Crossley, & Kyle, 2017a). Generally, more proficient learners are more likely to have access to words with longer latencies in L1 speakers. Berger et al. (2017a) have shown that word recognition norms can be used to model both longitudinal development and holistic scores of lexical proficiency.

Contextual Distinctiveness

From the usage-based perspective, the range of contexts in which a word is encountered can be a factor in determining the difficulty of the lexical item. Contextual distinctiveness is a

category of indices that aim to tap into this variation. The context here can be operationalized in three ways: (a) the range of human psychological association related to the word (Kiss et al., 1973), (b) the range of lexical contexts (McDonald co-occurence probability; McDonald & Shillcock, 2001), and (c) the range of semantic domains in which the word is used (semantic diversity [SemD]; Hoffman, Ralph, & Rogers, 2013). On average, a more proficient learner will produce words that are used in more restricted associative, lexical, and semantic contexts (Berger, Crossley, & Kyle, 2017b; Kyle et al., 2017).

Word Neighborhood

Sophisticated lexical items tend to have more complex formal features. Word neighborhood concerns how distinct a word is based on the phonological or orthographical features it shares with other words (Yarkoni, Balota, & Yap, 2008). Words with fewer orthographic and/or phonological neighbors are formally complex, thus processed more slowly (Andrews, 1989). They are more likely to occur in essays that earn higher word-choice scores (Kyle et al., 2017).

Academic Language

One of the approaches to defining sophisticated lexis is to tap into the extent to which a learner uses academic language in their texts, such as words in the academic word list (Coxhead, 2000) and phrases documented in the academic formulas list (Simpson–Vlach & Ellis, 2010). The use of academic language has been found to contribute to the quality of spoken as well as written production (Kim et al., 2018; Kyle & Crossley, 2015).

N-gram: Frequency, Range, and Association Strength

Many researchers argue that collocation should be included in the measures of lexical richness (Bestgen & Granger, 2014; Bulté & Housen, 2012; Nation & Webb, 2011) because it allows us to tap into the 'use' aspect of the vocabulary knowledge framework (Nation, 2013). In previous research, multiword units (e.g., collocations) were found to contribute to overall language performance (Boers et al., 2006) and to play an important part during speech production (Kormos, 2006). One typical way to identify multiword units is a form-based corpus approach wherein frequency (i.e., the number of occurrences), range (i.e., the number of documents), and association strength (see the next section) of n-grams (i.e., contiguous sequences of *n* words; O'Donnell, Römer, & Ellis, 2013; Paquot & Granger, 2012) are investigated.

In corpus linguistics, researchers use association measures to highlight n-grams with varying qualities (Evert, 2005). According to Gablasova, Brezina, and McEnery (2017), one of the distinguishing features of the association measures is exclusivity, which refers to the extent to which words in a given combination occur predominantly with each other. A common measure of this kind includes mutual information (MI), which highlights rare exclusivity by overemphasizing lower frequency combinations. A mathematical transformation of MI wherein the numerator is squared, called MI^2, can overcome this limitation of MI, and thus can be a more valid measure of exclusivity (Gablasova et al., 2017).[3]

CAPTURING TASK-SPECIFIC LEXICAL PERFORMANCES THROUGH SOPHISTICATION INDICES

Recent studies using the previously mentioned 12 categories of lexical sophistication indices have demonstrated that multiple index types are indeed linked to rater judgements. More importantly, the studies have suggested that the extended notion of lexical sophistication can be used to describe a set of specific lexical features associated with the genre or register of the assessment tasks, or target language use (Bachman & Palmer, 2010). For instance, Kyle et al. (2016) showed that, among others, range and meaningfulness can successfully differentiate performances in Test-of-English-as-a-Foreign-Language (TOEFL) independent speaking from those in integrated speaking tasks. Similarly, Kyle and Crossley (2016) reported that the indices can capture task-specific lexical features (e.g., range, hypernymy) between integrated and independent TOEFL Internet-based Test (iBT) writing tasks.

A review of existing literature further suggests that some index categories resulted in more stable associations with proficiency scores (broadly construed) across different domains of language use than other categories of indices (see Supporting Information A for a summary). Among the variety of categories, psycholinguistic information (e.g., concreteness) demonstrates consistent negative correlations with the different proficiency measures (e.g., Crossley et al., 2011a; Saito et al., 2016a). However, such stable relations have not been obtained with indices related to

frequency and semantic networks, especially in relation to spoken tasks (e.g., Kyle & Crossley, 2015; Saito et al., 2016a). This may indicate that factors related to elicitation tasks play a key role mediating the predictive relationships between the categories of indices and the proficiency scores.

Additionally, there is an insufficient amount of research to make generalizations about a number of new index types. These indices include word recognition norms, contextual distinctiveness, word neighbors, and n-gram association strengths (Kyle et al., 2017). They have been used to predict written, but not spoken, proficiency (though see Kim et al., 2018, for a longitudinal analysis of a spoken task). Further study is needed to investigate the extent to which these indices tap into lexical sophistication in different tasks. Further, due to the large number of (potentially overlapping) indices that are now available, it is useful to determine whether these hundreds of indices can be meaningfully reduced into a smaller number of latent dimensions of lexical use. Although one study has investigated this issue in a corpus of argumentative writing (Kim et al., 2018), no studies that we are aware of have done so in a spoken task.

Taken together, the current study attempts to investigate the dimensions of sophisticated lexical use in an underresearched register/genre: an OPI. An OPI is considered to provide a speaker-centered as well as here-and-now discourse (Staples, LaFlair, & Egbert, 2017), which situationally contrasts with the argumentative written task investigated in Kim et al. (2018). Based on the previous findings (see Supporting Information A), it is of interest whether and how a set of indices from the aforementioned 12 categories would jointly characterize sophisticated lexical use in the OPI.

Accordingly, this study is guided by the following research questions (RQs):

RQ1. What are the subdimensions of lexical sophistication in OPIs?
RQ2. Which of the identified subdimensions of lexical sophistication are predictive of the holistic scores of OPIs, and to what extent?

METHOD

To investigate the research questions mentioned, an EFA (to address RQ1) was conducted, followed by a stepwise multiple linear regression (to address RQ2).

Learner Corpus

The learner corpus selected for this study was the National Institute of Information and Communications Technology Japanese Learner English (NICT JLE) corpus, which includes 1,281 transcribed OPIs by Japanese learners of L2 English (Izumi, Uchimoto, & Isahara, 2004). The OPI used in the collection of this corpus was developed in collaboration with the American Council on the Teaching of Foreign Languages (ACTFL) to assess functional speaking ability (ACTFL-ALC Press, 1996). This version (ACTFL-ALC Standard Speaking Test [SST]) was designed to accommodate a less proficient target population than the original ACTFL OPI (corresponding to novice to advanced learners) through the inclusion of different tasks (ACTFL-ALC Press, 1996; Koizumi & Hirai, 2012). The SST comprises a series of speaking tasks that include self-introductions, single picture descriptions (e.g., neighborhood, restaurant), role plays (e.g., shopping, invitation), and cartoon descriptions (e.g., department store, car accident). The trained interviewer "informally evaluates the test-taker's level based on his/her responses and selects tasks appropriate to the level" (Koizumi & Hirai, 2012, p. 42).

The NICT JLE corpus provides the transcription of the recorded OPIs as well as the scores rated by two raters (or three if the initial two ratings disagreed) using a holistic rubric, which focuses on a range of aspects such as temporal aspects, grammar, and lexis (see Supporting Information B for the complete rubric). The descriptive statistics of the corpus are shown in Table 1. For the analysis of lexical sophistication, all the utterances by the interviewer were deleted from the transcription, and resulting transcriptions were analyzed with TAALES 2.8.

Selection of Lexical Sophistication Indices

In order to represent the construct of lexical sophistication, 12 categories that have been used in previous research (all reviewed in the Multidimensional Lexical Sophistication section) were selected from the indices calculated by TAALES 2.8 (Kyle et al., 2017). One advantage of TAALES is its comprehensiveness and transparency. Not only does it cover the ever-growing subdomains of lexical sophistication, it provides researchers with individual item outputs for qualitative manual inspection of the data (for more information and formulas for each index, visit https://www.linguisticanalysistools.org/taales.html). Since TAALES includes various related

TABLE 1
Overview of National Institute of Information and Communications Technology Japanese Learner English Corpus

	Number of Words			Score					
N	Total	Mean	*SD*	Min	Max	Mean	*SD*	Skewness	Kurtosis
1,281	1,102,192	860.415	327.418	1	9	4.664	1.574	.936	.550

Note. *SD* = standard deviation.

versions of each index (e.g., raw and logarithmically transformed frequencies; all-word, content-word only, function-word only indices, etc.), a two-step index selection process was developed in order to avoid the use of closely related indices. First, indices from each related group were refined to reduce conceptually overlapping indices. The remaining indices were then checked statistically for multicollinearity (see Analysis section), which can negatively affect multivariate analyses (Tabachnick & Fidell, 2013).

First of all, for frequency, range, and strength of association, only norms from spoken corpora were used in order to more closely match characteristics of the learner corpus. Given this criterion, frequency, range, and strength of association norms (as applicable) were drawn from G. Brown's (1984) list of frequency counts derived from the London-Lund Corpus (Svartvik & Quirk, 1980), the British National Corpus (BNC)–Spoken (BNC Consortium, 2007), SUBTLEXus (Brysbaert & New, 2009), and the spoken component of the Corpus of Contemporary American English (COCA; Davies, 2009). Second, only content-word (CW) and function-word (FW) indices were considered[4] (all-word indices were excluded) to reduce redundancy and provide a fine-grained approach. Third, nontransformed indices were selected over the transformed indices where applicable (i.e., where the nontransformed indices were normally distributed in the learner data). Finally, all the other indices that were inclusive of the others (e.g., word neighbors including homophones over those excluded) were chosen to reduce index redundancy. By employing these criteria, 110 out of the 132 indices from the 12 categories were selected (see Supporting Information C for a summary of these indices).

Analysis

RQ1: Exploratory Factor Analysis. Because large-scale, multifaceted studies of lexical sophistication are rare in the literature, an exploratory approach was taken with a goal of identifying latent factors of lexical sophistication (Plonsky & Gonulal, 2015; Sawaki, 2013). Accordingly, an EFA using the minimum residual method (Tabachnick & Fidell, 2013) was conducted using R 3.3.2 (R Development Core Team, 2014) with the psych package (Revelle, 2016), rather than simply reducing the number of variables. The minimum residual method was chosen for two reasons: First, the principal axis factor method (which has been commonly used in applied linguistics research) is less amenable for generalization since it assumes that the data is sampled entirely from the population (Field, 2009; Loewen & Gonulal, 2015). Second, the minimum residual is considered more tolerant to the violation of the multivariate normality assumption of the dataset than maximum likelihood (Zygmont & Smith, 2014).

Before conducting the EFA, the univariate normal distributions as well as the linearity of the randomly selected indices were first confirmed using the visualization options in the WEKA statistical package (Hall et al., 2009). Second, the remaining dataset was checked for multicollinearity. Any two variables that were correlated at $|r| \geq .900$ were flagged for further analysis (Tabachnick & Fidell, 2013). In each multicollinear pair, the following protocol was used to decide which index would remain in the study. First, COCA–Spoken indices were prioritized because they include the widest range of indices in TAALES (e.g., unigram frequency and range, n-gram frequency, range, and association strength). SUBTLEXus, which includes subtitles from American television series, was kept at the next level because it was considered to better reflect the previous exposure of the learners in the NICT JLE corpus than other two (i.e., BNC and Brown). This was followed by BNC–Spoken (which includes informal conversations in the United Kingdom), which was considered to better represent the previous exposure of the learners than Brown's verbal frequency list (which documents spoken British English in the 1960s–1970s). Note that this procedure was relevant only

when any indices showed multicollinearity, and all noncollinear indices were entered into the EFA (see Supporting Information C).

After controlling for multicollinearity, the factorability of the dataset was examined using the Kaiser–Meyer–Olkin (KMO) measure of sampling adequacy and Bartlett's Test of Sphericity to see if the sample size and correlations among the indices were sufficient for an EFA. In determining the number of factors to extract, multiple techniques such as eigenvalues, scree plot, and cumulative percentage explained were used (Loewen & Gonulal, 2015; Zygmont & Smith, 2014). To obtain a factor loading matrix, oblique rotation (e.g., promax and oblimin) was chosen because correlations among the factors (i.e., subconstructs of lexical sophistication) were assumed among the subdimensions. Subsequently, pattern matrices with both oblimin and promax rotation were examined for the interpretability of these factors. Once an interpretable solution was obtained, several types of poorly behaved indices were identified. Indices with weak loadings ($\lambda <$ +/− .300)[5] and low communalities ($h^2 < .300$), the estimated amount of total variance in the specific index that is explained by all the factors (T. A. Brown, 2015), were flagged and examined substantively in terms of the construct representativeness (T. A. Brown, 2015). Similarly, indices that lowered the reliability index (i.e., Cronbach's alpha) of each factor were marked as well. If these indices were not important in interpreting the factor, they were deleted from the model in order to produce a parsimonious solution. Meanwhile, instances of cross-loadings were retained if they were conceptually related to the multiple factors. Subsequently, EFA was iterated to ensure that the dataset still produced interpretable solutions (Loewen & Gonulal, 2015).

RQ2: Multiple Regression. To examine the extent to which the resulting factors predict the holistic score in the current OPI (RQ2), a multiple regression was conducted with the factor scores in EFA (the weighted sum scores) on the holistic scores of the OPI with R (R Development Core Team, 2014). Due to the exploratory nature of this study and the lack of precursor studies on multidimensional lexical sophistication in oral language, a stepwise regression was used. Prior to conducting the regression analysis, the factor scores were examined in terms of univariate normality and linearity of the relationships using histograms and scatterplots. Any factor scores that were not normally distributed or did not show meaningful relationships with the holistic score ($|r| \geq .100$) were discarded. Subsequently, the variables were examined in terms of the stricter criteria of multicollinearity for multiple regression (Tabachnick & Fidell, 2013). Among the sets of the collinear factors, the one with the strongest correlation with the OPI ratings was kept.

The resulting set of variables (i.e., composite scores by factor loading) were entered into a stepwise multiple linear regression using the Akaike information criterion (AIC) method (Akaike, 1974), which minimizes the deviation of the model from the dataset while using the fewest predictors. This was followed by a 10-fold cross-validated forced entry linear regression with the variables from the final stepwise regression model. In 10-fold cross-validation, the dataset is randomly divided into 10 folds (sections) of equal size. Nine folds are used as a training set to create a predictor model, which is then tested on the 10th fold. This procedure is repeated until each fold has been used as the testing set (Witten & Frank, 2005). Cross-validation enabled us to ensure that the predictor model was consistent across the dataset.

RESULTS

RQ1: Exploratory Factor Analysis

In order to investigate how each category of index works together to measure lexical sophistication in the OPI, an EFA was conducted. None of the indices violated the assumption of a normal distribution, but 19 of 110 indices were removed due to multicollinearity (for EFA; $|r| \geq .900$; see Supporting Information C for the removed indices), resulting in 91 remaining indices (KMO measure of sampling adequacy = .823; Bartlett's Test of Sphericity < .001). As a reference, Supporting Information D presents descriptive statistics and correlation coefficients between the 91 indices and the OPI holistic scores.

In the EFA, as a result of visual inspection of the scree plot, multiple solutions (i.e., 7–10 factor solutions) were considered initially. After examining the pattern matrices, a 10-factor solution with oblimin rotation was selected in terms of a clearer separation of the factor components and total variance explained. This model included a total of 13 poorly behaved indices, all of which were eliminated according to the preset criteria (see Supporting Information C for these indices). None of the cross-loaded indices were removed. Finally, EFA with 78 indices was rerun to obtain a more parsimonious solution, resulting in

TABLE 2
Eigenvalues and Cumulative Percentages of the Variance Explained by Factors

Factor	Eigenvalue	Proportion of Variance Explained	Cumulative Variance Explained	Portion Explained	Cumulative Portion Explained
F1: Abstract but common CWs	13.520	.104	.104	.158	.158
F2: CW neighborhood and access ease	9.053	.090	.195	.137	.295
F3: Frequent FWs	7.249	.075	.270	.114	.408
F4: Strongly associated, common trigrams	5.534	.065	.334	.098	.506
F5: Concrete FWs with frequent neighbors	3.487	.058	.392	.087	.594
F6: CW age of exposure/acquisition	2.828	.058	.450	.088	.681
F7: Strongly associated, common bigrams	2.100	.062	.512	.094	.775
F8: FW access difficulty	1.686	.050	.562	.076	.851
F9: FW neighborhood	1.080	.050	.611	.075	.926
F10: Exclusive collocations (MI)	.630	.049	.660	.074	1.000

Note. CW = content word; FW = function word; MI = mutual information.

TABLE 3
Interfactor Correlations of the 10 Factor Solution

Factor	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10
F1: Abstract but common CWs	1.000									
F2: CW neighborhood and access ease	.177	1.000								
F3: Frequent FWs	.071	.012	1.000							
F4: Strongly associated, common trigrams	.197	−.214	−.002	1.000						
F5: Concrete FWs with frequent neighbors	−.074	−.352	−.021	.056	1.000					
F6: CW age of exposure/acquisition	.125	.070	.044	−.071	.029	1.000				
F7: Strongly associated, common bigrams	.121	−.101	.267	.104	.005	.094	1.000			
F8: FW access difficulty	.354	.085	.179	.080	−.124	.453	.090	1.000		
F9: FW neighborhood	.110	−.001	.082	.024	−.093	−.031	.030	.206	1.000	
F10: Exclusive collocations (MI)	−.057	−.198	−.296	.082	.071	.263	−.070	.022	−.096	1.000

Note. CW = content word; FW = function word; MI = mutual information.

another 10-factor solution without any problematic indices (KMO = .837; see Table 2 for eigenvalues and variance explained in this model and Table 3 for the interfactor correlations). The full pattern matrix of the final solution can be found in Supporting Information E. Each factor (including loadings and communalities) is described as follows.

Factor 1: Abstract but Common Content Words. Factor 1 comprised a range of indices related to concreteness, contextual distinctiveness, word range, meaningfulness, and frequency. See Table 4 for the factor loadings for this component. Among the strongly loaded indices, concreteness and meaningfulness were negatively associated (−.884 and −.787), meaning that the factor was associated with less concrete and less meaningful CWs. On the other hand, the positive loadings of frequency, contextual distinctiveness, and range indices suggested that these were more

TABLE 4
Factor Loadings and Communalities of Each Index in Factor 1

Factor 1: Abstract but Common CWs ($\alpha = .939$)	Loading	h^2	α if Deleted
Brysbaert concreteness combined CW	−.884	.816	.931
Hoffman et al. semantic distinctiveness CW	.802	.745	.932
COCA–Spoken range CW	.792	.947	.927
MRC meaningfulness CW	−.787	.785	.933
BNC–Spoken frequency CW	.714	.798	.929
SUBTLEXus range CW	.679	.880	.931
Brown frequency CW	.651	.703	.933
COCA–Spoken frequency CW	.639	.763	.932
SUBTLEXus frequency CW	.548	.663	.936
Hypernymy nouns (sense 1, path 1)	−.534	.371	.944
COCA–Spoken bigram proportion 30k	.520	.848	.935
Hypernymy nouns (sense mean, path 1)	−.498	.301	.944

Note. CW = content words; COCA = Corpus of Contemporary American English; MRC = Medical Research Council Psycholinguistic Database; BNC = British National Corpus.

frequent, widely used words that have ambiguous meanings. Other weakly loaded indices such as hypernymy (−.534) suggested that the factor was related to more superordinate nouns than specific ones. Therefore, a text that earned a high abstract, but common CWs score (F1) would be characterized by less concrete but more frequent and widely used content words (e.g., *condition*, *chance*, *different*, *know*, *thing*).

Factor 2: Content Word Neighborhood and Access Ease. Factor 2 included CW neighborhood indices and word recognition indices (see Table 5). Overall, texts that earned a high score for CW neighborhood and access (F2) would be characterized by CWs that have high frequency orthographic and phonological neighbors as well as a higher number of orthographic and phonological neighbors, and that are processed more quickly by L1 speakers.

Factor 3: Frequent Function Words. Indices loaded onto Factor 3 included FW frequency, familiarity, neighborhood frequency, and other related indices (see Table 6). A text earning a high score on this factor would include FWs that are both frequent and familiar.

Factor 4: Strongly Associated, Common Trigrams. All the indices included in Factor 4 were related to trigram indices (see Table 7). Since T-score is considered as a measure of adjusted frequency (Gablasova et al., 2017), the stronger loadings of the range, T-score, and frequency indicated that a text that obtained a high strongly associated, common trigrams score (F4) included more commonly used trigrams.

Factor 5: Concrete Function Words With Frequent Neighbors. Factor 5 was characterized by FW indices related to concreteness, orthographic and phonological neighbor frequency (and distance), and AoA (see Table 8). A text that earned a high concrete FW with frequent neighbors score (F5) would therefore be characterized by FWs that are less conceptually abstract and that have less sophisticated neighbors.

Factor 6: Content Word Age of Exposure/Acquisition. Indices loaded onto Factor 6 primarily included AoE or AoA indices, CW familiarity, academic words, and contextual distinctiveness (see Table 9). In particular, AoE/A indices loaded more strongly onto the factor (.851–.587), followed by less familiar words (−.463). A text that earned a high CW AoE or AoA score (F6) would be characterized by CWs that are learned later and are less familiar.

Factor 7: Strongly Associated, Common Bigrams. Factor 7 mainly comprised bigram indices (see Table 10). The strongest loading onto this factor was MI^2 (.820), which highlights exclusive collocations (Gablasova et al., 2017), followed by range and other association measures (.726–.580). A text that obtained a high strongly associated, common bigrams score (F7) would be characterized by exclusive bigrams that are used in wider contexts.

Factor 8: Function Word Access Difficulty. Factor 8 was related positively with two response latencies in the recognition norm (i.e., lexical decision and word naming; .877 and .595) and negatively

TABLE 5
Factor Loadings and Communalities of Each Index in Factor 2

Factor 2: CW Neighborhood and Access Ease (α = .918)	Loading	h^2	α if Deleted
Average log hyperspace analogue to language (HAL) frequency of closest orthographic neighbors CW	.924	.877	.901
Average log hyperspace analogue to language (HAL) frequency of closest phonological neighbors CW	.876	.819	.904
Phonological neighbors CW (homonyms included)	.825	.791	.903
Orthographic neighbors CW	.785	.758	.904
Phonographic neighbors CW (homophones excluded)	.744	.584	.910
Average Levenshtein distance of closest phonological neighbors CW	−.735	.753	.907
Phonological neighborhood frequency CW (homophones included)	.674	.474	.915
Orthographic neighborhood frequency CW	.622	.470	.915
Phonographic neighborhood frequency logarithm CW (homophones included)	.469	.408	.921
Lexical decision time CW	−.418	.492	.916
Word naming response time CW	−.397	.504	.915
Word naming response time (standard deviation) CW	−.344	.273	.921

Note. CW = content word; *SD* = standard deviation.

TABLE 6
Factor Loadings and Communalities of Each Index in Factor 3

Factor 3: Frequent FWs (α = .879)	Loading	h^2	α if Deleted
BNC–Spoken frequency FW	.957	.980	.845
Brown frequency FW	.931	.916	.849
SUBTLEXus frequency FW	.830	.928	.857
MRC familiarity FW	.624	.620	.863
COCA–Spoken frequency FW	.614	.800	.858
Phonological neighborhood frequency (homophones included) FW (homophones included)	.480	.373	.869
Average log hyperspace analogue to language (HAL) frequency of closest phonological neighbors FW	.446	.512	.871
BNC–Spoken bigram frequency	.415	.602	.877
BNC–Spoken range FW	.387	.439	.883
Free association tokens response FW	.381	.183	.894

Note. FW = function word; BNC = British National Corpus; MRC = Medical Research Council Psycholinguistic Database; COCA = Corpus of Contemporary American English.

with accuracy of the norms (−.792 and −.627; see Table 11). A text with high FW access difficulty (F8) would thus be characterized by FWs that are more difficult to access.

Factor 9: Function Word Neighborhood. Factor 9 was strongly associated with FWs that have more neighbors and occur in wider contexts (i.e., contextual distinctiveness [.490] and range [.468 and .415]; see Table 12). A text that earned a high FW neighborhood score (F9) would be characterized by FWs with more orthographic and phonologic neighbors.

TABLE 7
Factor Loadings and Communalities of Each Index in Factor 4

Factor 4: Strongly Associated, Common Trigrams ($\alpha = .921$)	Loading	h^2	α if Deleted
COCA–Spoken trigram range	.988	.991	.901
COCA–Spoken trigram bigram to unigram association strength (T)	.822	.835	.898
COCA–Spoken trigram unigram to bigram association strength (T)	.763	.804	.901
BNC–Spoken trigram frequency	.731	.488	.920
COCA–Spoken trigram unigram to bigram association strength (DP)	.651	.538	.916
COCA–Spoken trigram bigram to unigram association strength (MI^2)	.570	.922	.902
COCA–Spoken trigram bigram to unigram association strength (DP)	.500	.603	.923

Note. COCA = Contemporary Corpus of American English; T = T-score; BNC = British National Corpus; DP = Delta P, operationalized as P(n-gram|Item1) – P(n-gram|Item1) (see Kyle et al., 2017, for more information), MI^2 = mutual information with squared numerator.

TABLE 8
Factor Loadings and Communalities of Each Index in Factor 5

Factor 5: Concrete FWs With Frequent Neighbors ($\alpha = .861$)	Loading	h^2	α if Deleted
Brysbaert concreteness combined FW	.721	.782	.862
Orthographic neighborhood frequency FW	.709	.756	.817
Average Levenshtein distance of closest phonological neighbors FW	−.702	.776	.845
Average log hyperspace analogue to language (HAL) frequency of closest orthographic neighbors FW	.702	.751	.815
Phonographic neighborhood frequency logarithm (homophones included) FW (homophones included)	.688	.615	.837
Age of acquisition FW	−.641	.672	.847

Note. FWs = function words.

TABLE 9
Factor Loadings and Communalities of Each Index in Factor 6

Factor 6: CW Age of Exposure/Acquisition ($\alpha = .840$)	Loading	h^2	α if Deleted
Latent Dirichlet allocation age of exposure (.40 cosine threshold)	.851	.754	.795
Latent Dirichlet allocation age of exposure (inverse average)	.842	.768	.792
Latent Dirichlet allocation age of exposure (inverse slope)	.769	.573	.811
Age of acquisition CW	.587	.699	.809
MRC familiarity CW	−.463	.322	.828
Academic word list all	.437	.258	.831
McDonald co-occurrence probability CW	.417	.369	.851
Free association stimuli elicited CW	−.408	.389	.842

Note. CW = content word; MRC = Medical Research Council Psycholinguistic Database.

Factor 10: Exclusive Collocations. This factor (see Table 13) included three indices of mutual information (MI) for both bigrams and trigrams (.916–.704). The divergence of the MI indices from the other n-grams, especially from MI^2 (as seen in Tables 10 and 12), implied that this factor captured either highly exclusive n-grams or those composed of infrequent words (Gablasova et al., 2017). Therefore, it was clear from this contrast that a text that earned a high exclusive collocations score (F10) would be characterized by the use of highly exclusive n-grams.

TABLE 10
Factor Loadings and Communalities of Each Index in Factor 7

Factor 7: Strongly Associated, Common Bigrams ($\alpha = .911$)	Loading	h^2	α if Deleted
COCA–Spoken bigram association strength (MI^2)	.820	.926	.874
COCA–Spoken bigram range	.726	.956	.878
COCA–Spoken bigram association strength (T)	.705	.750	.894
COCA–Spoken bigram association strength (AC)	.651	.812	.882
COCA–Spoken bigram association strength (DP)	.580	.570	.931
COCA–Spoken trigram proportion 30k	.515	.884	.906

Note. AC = approximate collexeme strength, T = T-score, DP = Delta P, MI^2 = mutual information with squared numerator.

TABLE 11
Factor Loadings and Communalities of Each Index in Factor 8

Factor 8: FW Access Difficulty ($\alpha = .817$)	Loading	h^2	α if Deleted
Lexical decision time FW	.877	.828	.754
Lexical decision accuracy FW	−.792	.693	.774
Word naming response accuracy FW	−.627	.479	.799
McDonald co-occurrence probability FW	.598	.442	.812
Word naming response time FW	.595	.441	.777
SD lexical decision time FW	.548	.525	.798
SD word naming response time FW	.411	.330	.825

Note. FW = function word; *SD* = standard deviation.

TABLE 12
Factor Loadings and Communalities of Each Index in Factor 9

Factor 9: FW Neighborhood ($\alpha = .799$)	Loading	h^2	α if Deleted
Orthographic neighbors FW	.831	.829	.741
Average Levenshtein distance of closest orthographic neighbors FW	−.824	.782	.728
Phonographic neighbors FW (homophones excluded)	.553	.571	.794
Hoffman et al. semantic distinctiveness FW	.490	.554	.778
COCA–Spoken range FW	.468	.782	.763
Phonological neighbors FW (includes homonyms)	.441	.629	.813
SUBTLEXus range FW	.415	.418	.783

Note. FW = function word; COCA = Corpus of Contemporary American English.

TABLE 13
Factor Loadings and Communalities of Each Index in Factor 10

Factor 10: Exclusive Collocations (MI) ($\alpha = .892$)	Loading	h^2	α if Deleted
COCA–Spoken trigram bigram to unigram association strength (MI)	.916	.935	.804
COCA–Spoken trigram unigram to bigram association strength (MI)	.854	.778	.806
COCA–Spoken bigram association strength (MI)	.704	.738	.921

Note. MI = mutual information; COCA = Corpus of Contemporary American English.

RQ2: Modeling Oral Proficiency Interview Scores

To investigate the relationship between the latent factors and holistic OPI scores in the NICT JLE corpus, a multiple linear regression was conducted using the factor scores to predict holistic OPI scores. All of the 10 factors demonstrated normal distributions. Two of the factors—namely, CW neighborhood and access ease (F2) and CW AoE/AoA (F6)—did not demonstrate

TABLE 14
Correlations Between Holistic Essay Score and Lexical Sophistication Factors

Factor Name	Correlation With Holistic Score
F1: Abstract but common CWs	.625
F7: Strongly associated, common bigrams	.492
F5: Concrete FWs with frequent neighbors	−.414
F9: FW neighborhood	.277
F8: FW access difficulty	.203
F10: Exclusive collocations (MI)	.168
F3: Frequent FWs	.151
F4: Strongly associated, common trigrams	.143
F2: CW neighborhood and access ease[a]	−.071
F6: CW age of exposure/acquisition[a]	−.036

Note. CW = content word; FW = function word; MI = mutual information.
[1]Factor excluded from regression analysis due to weak correlation ($|r| < .1$).

TABLE 15
Summary of Multiple Regression Model

Entry	Predictors Included	r	Adjusted R^2	R^2 Change	B	SE	β
1	F1: Abstract but common CWs	.625	.391	.391	.105	.006	.424
2	F3: Frequent FWs	.628	.395	.004	.048	.008	.137
3	F5: Concrete FWs with frequent neighbors	.702	.493	.098	−.166	.010	−.337
4	F7: Strongly associated, common bigrams	.724	.524	.031	.046	.011	.099
5	F8: FW access difficulty	.730	.534	.010	.061	.009	.123
6	F9: FW neighborhood	.748	.559	.025	.107	.011	.192
7	F10: Exclusive collocations (MI)	.758	.575	.016	.106	.015	.152

Note. Estimated constant term = 4.663, B = unstandardized beta, SE = standard error; β = standardized beta. CW = content word; FW = function word; MI = mutual information.

a meaningful correlation ($|r| \geq .100$) with the holistic score and were removed from the subsequent analysis (see Table 14). None of the remaining indices demonstrated multicollinearity. Therefore, the remaining eight factors were entered into a stepwise multiple regression. The resulting model, which included seven factors, accounted for 57.5% ($r = .758$, adjusted $R^2 = .575$) of the total variance in OPI score (see Table 15 for the model). The 10-fold cross-validation revealed that the current model explained 57.58% of the variance ($R^2 = .576$), suggesting that the model explored previously is consistent across the dataset (See Supporting Information F for example excerpts illustrating the relationships between lexical sophistication and OPI performance).

DISCUSSION

Recent developments in lexical sophistication have been promising in identifying various lexical features mediated by both proficiency and task designs (Kyle & Crossley, 2016; Kyle et al., 2016); however, previous studies have extensively focused on certain types of assessment situations (e.g., argumentative writing). Therefore, it was still unclear to what extent categories-of-sophistication indices represent separate subdimensions in different contexts, especially spoken mode (Kim et al., 2018). Accordingly, the current study explored subdimensions of lexical sophistication in an OPI (ACTFL-ALC Press, 1996), which may influence the structure by providing a learner-centered, here-and-now discourse (Staples et al., 2017).

RQ1: Subdimensions of Lexical Sophistication

The first RQ was intended to uncover the number of lexical sophistication dimensions as well as their latent structure in an OPI task. Through an EFA, 10 subdimensions were identified in the current OPI data. A clear pattern that emerged was the separation of indices related to CWs, function words FWs, and n-grams. Three factors were

related to CW use (F1, F2, F6), four factors were related to FW use (F3, F5, F8, F9), and three factors were related to n-gram use (F4, F7, F10). This indicates that each unit of analysis contributes to an overall understanding of lexical use in OPIs by complementing with one another (see interfactor correlations such as between F1 and F8 [$r = .354$], and F1 and F4 [$r = .197$]).

With regard to the grouping of the lexical sophistication indices, we were interested in whether and the extent to which conceptually and operationally similar and/or diverse indices are related in the OPI. The EFA revealed that that five factors (F1, F2, F3, F5, and F6) included a diversity of indices in each while the other five factors (F4, F7, F8, F9, and F10) represented operationally similar ones.

The abstract but common CW factor (F1) included indices from various preset categories: concreteness, range, contextual distinctiveness, meaningfulness, frequency, and hypernymy. The current finding may indicate that these distinct index types are likely to measure co-occurrent lexical features in the OPI (Norris & Ortega, 2009). For some indices, this co-occurrence is easily interpreted because indices such as range and contextual distinctiveness are conceptually related despite the distinct operationalizations. Range indices simply count the number of documents in which the target word occurs, while contextual distinctiveness (as measured by SemD) measures the range of semantic contexts in which a word occurs using latent semantic analysis (Hoffman et al., 2013). The result may indicate that, despite the operational differences, the two indices are similar enough to be represented in a single dimension. On the other hand, the same factor also represented indices that are conceptually distinct—namely, frequency and psycholinguistic norms (e.g., concreteness, imageability). The two are considered separate categories in that frequent words are not necessarily concrete nor imageable (e.g., *case*, *time*, *get*) or vice versa (e.g., *pineapple*). With regard to learning, frequency is undoubtedly one of the most important driving forces of acquisition (Ellis, 2002), while psycholinguistic norms have been shown to be related to saliency, another important construct in language learning; as more concrete and imageable words are more salient in discourse and hence more noticeable (Crossley, Kyle, & Salsbury, 2016; Ellis & Beaton, 1993). All in all, the inclusion of the different categories of indices in F1 may suggest that this factor represents important task-specific lexical features that were observed in the current OPI data.

The other four subdimensions included conceptually related indices. For instance, CW neighborhood and access factor (F2) included indices that are conceptually related but operationally distinct such as those related to formal distinctiveness (e.g., phonological neighbors) and lexical access (e.g., word naming response time). These indices may capture a shared, underlying construct of lexical sophistication because words that have more neighbors tend to be more common and thus likely to be accessed more easily. In contrast, the remaining five factors (F4, F7, F8, F9, F10) comprised operationally related indices. The factor FW access difficulty (F8), for example, primarily included indices derived from word recognition norms (Berger et al., 2017a). Similarly, FW neighborhood (F9) was comprised mainly of neighborhood indices.

Two striking patterns were identified in relation to n-gram use (F4, F7, and F10). First, two factors (strongly associated, common trigrams [F4] and strongly associated, common bigrams [F7]) were sharply divided with regard to *n* length regardless of the operational differences (i.e., range, frequency, and association strength). This mirrors a corpus-based investigation of formulaic sequences among learner and professional discourses (O'Donnell et al., 2013), which demonstrated that (a) longer n-grams are less likely to frequently recur across corpora, and (b) the discrepancies between the shorter and longer sequences are more noticeable in learner corpora. The current finding would warrant further research focusing on the interface of use and acquisition of multiword units with varying lengths. Another important finding was the independence of MI indices (F10) from other n-gram indices (F4 and F7). This is possibly due to the nature of MI scores, which overemphasize n-grams that are composed of infrequent lexical items (Evert, 2005; Gablasova et al., 2017). The current divergence of MI especially from MI^2 appears to support that MI taps into multiword units with distinct qualities in L2 oral data.

These overall patterns can be compared with those in Kim et al. (2018), though important caveats are in order due to methodological differences across the two studies. One of the convergent findings is the separation of CWs, FWs, and n-gram indices. Both studies clearly differentiated these lexical units and demonstrated important contributions of each to L2 proficiency (see the next section). An important divergence concerns the amount of variance explained by particular factors or components across the two studies. In the current study, for example, the CW-related

index that explained the most variance was the abstract but common CW factor (F1; see Table 2), whereas in Kim et al., it was word acquisition properties, which comprised range, word neighbors, frequency, word recognition norms, AoE, and academic words (see Kim et al., 2018, Table 3; p.129). Although some common indices are found across these factors or components (e.g., range), other indices clearly contrast with each other, which may indicate that they represent task-specific lexical features. The inclusion of word recognition norms, AoE, and academic words in Kim et al. may have reflected the variance in the argumentative writing they investigated, whereas psycholinguistic norms and hypernymy noun indices may capture here-and-now discourse of the OPI (Staples et al., 2017). Future studies would need to confirm this finding by including comparable corpora in a single common factor model (e.g., EFA and confirmatory factor analysis). Now, we turn to the relationship between the subdimensions and the OPI scores in the following section.

RQ2: Modeling Oral Proficiency Interview Score

The results of the stepwise multiple regression (see Table 15) indicated that 57.5% of the holistic OPI score was explained with seven factors: one related to CWs, two from n-gram, and the other four from FWs. The largest amount of the variance (39.1%) was explained by the abstract but common CWs factor (F1). The positive relationship between F1 and the OPI score indicates that more proficient learners (as measured by the OPI rubric; Supporting Information B) will produce CWs that are more frequent in reference corpora but that are more abstract and contextually diverse. For psycholinguistic norms, for instance, the negative relationship with OPI performance adds further evidence that proficient learners would be able to produce less salient lexical items (Crossley et al., 2011a, 2016; Crossley & Skalicky, 2017). With regard to frequency, on the other hand, the positive association between frequency and the OPI score contrasts with traditional assumptions of this index (Laufer & Nation, 1995; Michel, 2017), but concur with the findings of Kyle and Crossley (2015), who demonstrated that more frequent lexical items were indicative of spoken proficiency (i.e., independent speech tasks in TOEFL iBT). As already mentioned, this would imply that the relationships between some lexical sophistication indices and proficiency measures may be mediated by task-related factors (Bachman & Palmer, 2010). This merits further investigation into the possible impacts of more specific task variables (e.g., task types, complexity) on lexical sophistication indices in future studies (see Alexopoulou et al., 2017, as an example study of this type).

Tentatively, the relationships can be explained by possible task requirements and rater expectations related to lexis in the respective assessment situations. In argumentative and source-based writings, use of less frequent vocabulary might be encouraged, likely because the tasks and rubrics alike are designed to assess range of vocabulary and word choice as parts of the construct (e.g., TOEFL independent writing rubrics; https://www.ets.org/s/toefl/pdf/toefl_writing_rubrics.pdf). In contrast, in spoken tasks that have been investigated so far (Kyle & Crossley, 2015), it was automaticity and effectiveness of lexis that were at the center of the construct (for the rubrics see https://www.ets.org/s/toefl/pdf/toefl_speaking_rubrics.pdf). This possibly minimized the impact of the 'rareness' of vocabulary in the previous studies. Similarly, a closer look at the rubric of the current OPI may support this view because one of the key constructs measured is the ability to respond to general topics, as opposed to the ability to elaborate ideas using more complex or sophisticated linguistic items (see Supporting Information B for the OPI rubrics).

The second largest amount of the variance (9.8%) was explained by the concrete FWs with frequent neighbors factor (F5). The relationships with OPI scores suggest that more proficient learners will use FWs that are more sophisticated (following a traditional interpretation). While three other FW factors (F3, F7, and F9) added a relatively small (3.9%) amount of the variance explained by the model, the small to moderate relationships with speaking proficiency, and the nature of their use diverged from the CW factors, suggest the importance of considering FW use. This is of particular importance given the common practice of either combining CW and FW use into a single category or ignoring FW use altogether (Crossley, Cobb, & McNamara, 2013; Laufer & Nation, 1995).

In addition to the individual words, the regression model showed the contribution of multiword units (F7: bigrams, and F10: highly exclusive collocations) to L2 performance (4.7%). More proficient OPI performance can be explained by the use of bigrams that are more exclusive (MI^2) and those used in wider contexts (range and T-score), as indicated by F7 (strongly associated, common bigrams; 3.1%). Furthermore, it is noteworthy that the score can further be explained by the

use of highly exclusive collocations (F10), as measured with the MI indices (Gablasova et al., 2017). Although the amount of variance explained is relatively small, the results should be interpreted to suggest that collocation use plays a unique role in determining the quality of OPIs (Bestgen & Granger, 2014; Kyle & Crossley, 2015). Formulaic sequences, including n-grams, are generally considered to facilitate fluent, appropriate language performance in a given communicative context (Boers et al., 2006; Wray, 2002); the current result support the idea that collocational use should be added to the overall construct of lexical sophistication, the use of 'advanced' vocabulary items (Bulté & Housen, 2012; Crossley et al., 2015; Nation & Webb, 2011).

Implications

The results of the current study have implications with regard to the measurement of lexical sophistication as well as development of vocabulary use. First, the results confirmed the idea that lexical sophistication—"unusual" or "advanced" vocabulary use (Read, 2000, p. 203)–can be operationalized multidimensionally under the notions of relative complexity (Bulté & Housen, 2012) and intralexical difficulty (Ellis & Beaton, 1993; Laufer, 1997). Under the current framework, lexical sophistication can be redefined as a multidimensional concept that comprises not only (a) rareness (e.g., infrequent, narrower range) but also (b) conceptual features (e.g., less concrete, imageable), (c) formal distinctiveness (i.e., having fewer orthographic/phonological neighbors), (d) accessibility (i.e., requiring longer time to access), and (e) association strengths of the multiword units (i.e., stronger associations; see also Kim et al., 2018). Researchers will benefit from conceptualizing 'sophisticated' vocabulary use from the multidimensional perspective (see the 12 categories introduced in the literature review).

However, this does not mean that the 12 preexisting operationalized categories (e.g., range, contextual distinctiveness) measure completely separate aspects of sophistication across assessment tasks. Rather, as the current study demonstrates through a data-driven approach (i.e., EFA), more or less conceptually and/or operationally (dis)similar indices may co-represent patterns of lexical sophistication in a specific context of language use (see F1, F2, F3, F5, and F6 in this study).

Measuring lexical sophistication as a multidimensional construct also allows for register effects to be explored in more nuanced ways. Under a univariate frequency-based definition of lexical sophistication, register differences have sometimes been identified as measurement obstacles, particularly because certain types of discourses or topics (e.g., oral language) may not require infrequent words to be produced (Milton, 2009; Nation & Webb, 2011; Read, 2000). Since frequency was the only operationalized construct of sophistication in earlier research, a task had to be selected so that it would elicit enough 'infrequent' vocabulary items. In contrast, the multidimensionally operationalized lexical sophistication allows researchers to identify additional characteristics of 'advanced' vocabulary which may be sensitive to particular contexts of language use. The current study, for instance, demonstrated that lexical sophistication in the OPI can be characterized in terms of lexical features such as frequent but abstract word use and more strongly associated collocations (for analyses of expected vocabulary use in TOEFL iBT, see Kyle & Crossley, 2016; Kyle et al., 2016). Such a nuanced description of lexical use may not be possible in the previous notion of lexical sophistication.

In short, the current study suggests that approaching lexical sophistication multidimensionally would potentially uncover the nuanced characteristics of 'advanced' vocabulary use mediated by the functional requirements of the immediate tasks. Future studies would benefit from comparing and contrasting the nature of lexical sophistication associated with different contexts of language use (i.e., genre or register) to gain a deeper understanding of the construct and to shed light on the development and/or proficiency in lexical use.

LIMITATIONS AND FUTURE STUDIES

While the current study contributes to our understanding of the multidimensional nature of lexical sophistication, it also brings light to several methodological limitations and issues for further research. First, although the current study demonstrated potential task effects on lexical sophistication indices, the findings should be taken with caution because no direct comparison among different tasks was performed. Still, the review of the previous studies and the fact that the current results are in line with the previously found relationships among different tasks and major lexical sophistication indices (e.g., concreteness) would support that the results can still be linked to the task differences. However, we have to admit that this is indeed a tentative finding, and future research should replicate

the relationships with more directly comparable study designs. Second, due to the nature of the current OPI tasks, the current study did not investigate the possible topic–prompt effects. As lexical performance can be dependent on such factors, future research should investigate corpora including various tasks that can be comparable in terms of topics and prompts (Alexopoulou et al., 2017). Third, as one of the anonymous reviewers pointed out, lexical production in the current corpus could have been influenced by the dialogic nature of the OPI as learners may have incorporated some words from the examiner. Although this might be the case for natural conversations and has been a target of analysis in some studies (Crossley et al., 2016), the examiners in the current OPI rarely provide sophisticated vocabularies that might influence test takers' performances (see Supporting Information F for example excerpts). Last, since the current study investigated cross-sectional data, it confined the scope to the nature of the construct and possible connections to L2 learning. Future studies should validate the current view in the longitudinal design to illuminate how learners develop each dimension as a function of time or proficiency increase.

CONCLUSION

This study investigated the subdimensions of lexical sophistication in relation to OPIs, and explored the relationship between these dimensions and the proficiency scores. Through an EFA, 10 subdimensions of lexical sophistication were identified in OPI responses, where the factors represented each unit of analysis (i.e., CW, FW, and multiword units). Interestingly, 5 out of 10 factors comprised conceptually distinct indices, which are considered to jointly represent task-specific lexical sophistication (e.g., range, concreteness, contextual distinctiveness, hypernymy). Furthermore, contrary to the traditional assumption of the index, frequent words were found as a hallmark of a proficient OPI, mirroring some task-dependent behavior of frequency index. Taken together, this study provides further evidence of the multidimensional nature of the construct of lexical sophistication; suggests that while some features of lexical sophistication may be domain general, others may be task specific; and provides a roadmap for the measurement of lexical sophistication in future studies of proficiency, development, and use. Finally, we would like to encourage readers to refer to the Supporting Information, which contains summaries of previous studies, a complete list of indices, complete factor matrix, and example excerpts we were not able to include in the authors'.

ACKNOWLEDGMENTS

The authors are grateful to all anonymous reviewers and the journal editor, Professor Marta Antón, for constructive feedback on the earlier versions of the current article. The authors would also like to thank Professor Tetsuo Harada, Professor Yasuyo Sawaki, Takumi Uchihara, and Shungo Suzuki for providing critical comments and analytical suggestions to the earlier version of the draft. Their comments were helpful in improving the quality of the current article. Needless to say, all remaining errors are the authors'.

NOTES

[1] Only one of the seven subdimensions of lexical diversity proposed by Jarvis (2013)—lexical importance (also referred to as specialness or rarity)—is concerned with the kind of vocabulary produced by writers/speakers. The assumption, however, is that the choice of rare words may make greater contribution to human judgements of lexical diversity. Thus, this particular aspect of diversity may somewhat overlap with lexical sophistication at the level of operationalized measures. However, lexical diversity and lexical sophistication are still considered as distinct constructs.

[2] As one of the anonymous reviewers pointed out, the learning burden of lexical items includes both intra- and interlexical factors (Laufer, 1997). While intralexical factors concern difficulty features inherent to the item (e.g., frequency, concreteness, number of senses a word has), interlexical difficulty considers the interplay of lexical and learner factors (e.g., congruency of L1–L2 translation). Although we acknowledge the importance of interlexical difficulty, it was out of scope of the current study because (a) we are yet to operationalize the interlexical factors in a large-scale corpus study (e.g., availability of corpora and databases), and (b) such indices can inevitably become very specific to a narrower population.

[3] MI score is calculated as $MI = log_2 \frac{O}{\frac{R1*C1}{N}}$, where O denotes frequency of the combination, $R1$ denotes frequency of the first item, $C1$ denotes frequency of the second item, and N represents the total number of words in corpus. In this current formula, the ratio is inflated if the O, $R1$, and/or $C1$ is a small number. If O is squared as in MI^2, it will adjust the ratio between the observed and expected frequencies, and strike the balance between the frequent and infrequent multiword units (for more details see Evert, 2005; Gablasova et al., 2017).

[4] In TAALES 2.8, content words are operationally defined as nouns, adjectives, adverbs derived from

adjectives, and lexical verbs. All other words are considered function words.

[5] As recommended in the literature (T.A. Brown, 2015), the threshold loading value for the current EFA was determined based on the previous studies such as Staples et al. (2017).

Open Research Badges

This article has an Open Data Badge. Data available at: https://osf.io/jdyxf/?view_only=4f793abe052b4389b90fcb56fba7ce66.

REFERENCES

ACTFL-ALC Press. (1996). *Standard speaking test manual.* Tokyo: ACTFL-ALC Press.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716–723.

Alexopoulou, T., Michel, M., Murakami, A., & Meurers, D. (2017). Task effects on linguistic complexity and accuracy: A large-scale learner corpus analysis employing natural language processing techniques. *Language Learning*, *67*, 180–208.

Andrews, S. (1989). Frequency and neighborhood effects on lexical access: Activation or search? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*, 802–814.

Bachman, L. F., & Palmer, A. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world.* Oxford: Oxford University Press.

Berger, C. M., Crossley, S. A., & Kyle, K. (2017a). Using native-speaker psycholinguistic norms to predict lexical proficiency and development in second-language production. *Applied Linguistics*, *40*, 1–22.

Berger, C. M., Crossley, S. A., & Kyle, K. (2017b). Using novel word context measures to predict human ratings of lexical proficiency. *Educational Technology & Society*, *20*, 201–212.

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

BNC Consortium. (2007). *The British National Corpus* (version 3 [BNC XML ed.]). Accessed 5 July 2017 at http://www.natcorp.ox.ac.uk/

Boers, F., Eyckmans, J., Kappel, J., Stengers, H., & Demecheleer, M. (2006). Formulaic sequences and perceived oral proficiency: Putting a lexical approach to the test. *Language Teaching Research*, *10*, 245–261.

Brown, G. (1984). A frequency count of 190,000 words in the London-Lund corpus of English conversation. *Behavior Research Methods, Instruments, & Computers*, *16*, 502–532.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). New York: Guilford.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*, 977–990.

Bulté, B., & Housen, A. (2012). Defining and operationalising L2 complexity. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (pp. 21–46). Amsterdam: John Benjamins.

Coltheart, M. (1981). The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology Section A*, *33*, 497–505.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*, 213–238.

Crossley, S. A., Cobb, T., & McNamara, D. S. (2013). Comparing count-based and band-based indices of word frequency: Implications for active vocabulary research and pedagogical applications. *System*, *41*, 965–981.

Crossley, S. A., Kyle, K., & Salsbury, T. (2016). A usage-based investigation of L2 lexical acquisition: The role of input and output. *Modern Language Journal*, *100*, 702–715.

Crossley, S. A., Salsbury, T., & McNamara, D. (2009). Measuring L2 lexical growth using hypernymic relationships. *Language Learning*, *59*, 307–334.

Crossley, S. A., Salsbury, T., & McNamara, D. (2010). The development of polysemy and frequency use in English second language speakers. *Language Learning*, *60*, 573–605.

Crossley, S. A., Salsbury, T., & McNamara, D. S. (2015). Assessing lexical proficiency using analytic ratings: A case for collocation accuracy. *Applied Linguistics*, *36*, 570–590.

Crossley, S. A., Salsbury, T., McNamara, D. S., & Jarvis, S. (2011a). What is lexical proficiency? Some answers from computational models of speech data. *TESOL Quarterly*, *45*, 182–193.

Crossley, S. A., Salsbury, T., McNamara, D. S., & Jarvis, S. (2011b). Predicting lexical proficiency in language learner texts using computational indices. *Language Testing*, *28*, 561–580.

Crossley, S. A., & Skalicky, S. (2017). Examining lexical development in second language learners: An approximate replication of Salsbury, Crossley & McNamara (2011). *Language Teaching*, *29*, 1–21.

Daller, H., Milton, J., & Treffers–Daller, J. (2007). *Modelling and assessing vocabulary knowledge.* Cambridge: Cambridge University Press.

Davies, M. (2009). The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics*, *14*, 159–190.

de Jong, N. H., Steinel, M. P., Florijn, A. F., Schoonen, R., & Hulstijn, J. H. (2012). Facets of speaking proficiency. *Studies in Second Language Acquisition*, *34*, 5–34.

Ellis, N. C. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition*, *24*, 143–188.

Ellis, N. C., & Beaton, A. (1993). Psycholinguistic determinants of foreign language vocabulary learning. *Language Learning*, *43*, 559–617.

Evert, S. (2005). *The statistics of word cooccurrences word pairs and collocations* (Unpublished doctoral dissertation). Institut Fur Maschinelle Sprachverarbeitung Universitat Stuttgart, Stuttgart, Germany.

Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). London: SAGE Publications.

Gablasova, D., Brezina, V., & Mcenery, T. (2017). Collocations in corpus-based language learning research: Identifying, comparing, and interpreting the evidence. *Language Learning*, *67*(S1), 1–25.

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

Hall, M., Framk, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software. *ACM SIGKDD Explorations Newsletter*, *11*, 10–18.

Hoffman, P., Ralph, M. A. L., & Rogers, T. T. (2013). Semantic diversity: A measure of semantic ambiguity based on variability in the contextual usage of words. *Behavior Research Methods*, *45*, 718–730.

Izumi, E., Uchimoto, K., & Isahara, H. (2004). The NICT JLE Corpus: Exploiting the language learners' speech database for research and education. *International Journal of the Computer, the Internet and Management*, *12*, 119–125.

Jarvis, S. (2013). Capturing the diversity in lexical diversity. *Language Learning*, *63*, 87–106.

Kim, M. M., Crossley, S. A., & Kyle, K. (2018). Lexical sophistication as a multidimensional phenomenon: Relations to second language lexical proficiency, development, and writing quality. *Modern Language Journal*, *102*, 120–141.

Kiss, G. R., Armstrong, C., Milroy, R., & Piper, J. (1973). An associative thesaurus of English and its computer analysis. In A. J.Aitkin, R. W.Bailey, & N.Hamilton–Smith (Eds.), *The computer and literary studies* (pp. 153–165). Edinburgh: Edinburgh University Press.

Koizumi, R. (2012). Vocabulary and speaking. In C. A.Chapelle (Ed.), *The encyclopedia of applied linguistics*. Boston: John Wiley & Sons. https://doi.org/10.1002/9781405198431.wbeal1431

Koizumi, R., & Hirai, A. (2012). Comparing the story retelling speaking test with other speaking tests. *JALT Journal*, *34*, 35–60.

Koizumi, R., & In'nami, Y. (2013). Vocabulary knowledge and speaking proficiency among second language learners from novice to intermediate levels. *Journal of Language Teaching and Research*, *4*, 900–913.

Kormos, J. (2006). *Speech production and second language acquisition*. Mahwah, NJ: Lawrence Erlbaum.

Kuperman, V., Stadthagen–Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, *44*, 978–990.

Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly*, *49*, 757–786.

Kyle, K., & Crossley, S. A. (2016). The relationship between lexical sophistication and independent and source-based writing. *Journal of Second Language Writing*, *34*, 12–24.

Kyle, K., Crossley, S. A., & Berger, C. (2017). The tool for the automatic analysis of lexical sophistication (TAALES): Version 2.0. *Behavior Research Methods*, *50*, 1030–1046.

Kyle, K., Crossley, S. A., & McNamara, D. S. (2016). Construct validity in TOEFL iBT speaking tasks: Insights from natural language processing. *Language Testing*, *33*, 319–340.

Laufer, B. (1997). What's in a word that makes it hard or easy: Some intralexical factors that affect the learning of words. In N.Schmitt & M.McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy* (pp. 140–155). Cambridge: Cambridge University Press.

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical density in FL written production. *Applied Linguistics*, *16*, 307–322.

Loewen, S., & Gonulal, T. (2015). Exploratory factor analysis and principal components analysis. In L.Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 182–212). New York: Routledge.

Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *Modern Language Journal*, *96*, 190–208.

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, *42*, 381–92.

McDonald, S. A., & Shillcock, R. C. (2001). Rethinking the word frequency effect: The neglected role of distributional information in lexical processing. *Language and Speech*, *44*, 295–322.

Meara, P. (2005). Designing vocabulary tests for English, Spanish and other languages. In S. C.Butler, M. de los Á.Gómez–González, & S. M.Doval–Suárez (Eds.), *The dynamics of language use* (pp. 271–286). Amsterdam: John Benjamins.

Meara, P., & Bell, H. (2001). P_Lex: A simple and effective way of describing the lexical characteristics of short L2 texts. *Prospect*, *16*, 5–19.

Michel, M. (2017). Complexity, accuracy and fluency in L2 production. In S.Loewen & M.Sato (Eds.), *The handbook of instructed second language acquisition* (pp. 50–68). New York: Routledge.

Milton, J. (2009). *Measuring second language vocabulary acquisition*. Bristol, UK: Multilingual Matters.

Nation, P. (2013). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Nation, P., & Webb, S. (2011). *Researching and analyzing vocabulary*. Boston: Heinle Cengage Learning.

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

O'Donnell, M. B., Römer, U., & Ellis, N. C. (2013). The development of formulaic sequences in first and second language writing. *International Journal of Corpus Linguistics*, *18*, 83–108.

Paquot, M., & Granger, S. (2012). Formulaic language in learner corpora. *Annual Review of Applied Linguistics*, *32*, 130–149.

Plonsky, L., & Gonulal, T. (2015). Methodological synthesis in quantitative L2 research: A review of reviews and a case study of exploratory factor analysis. *Language Learning*, *65*(S1), 9–36.

R Development Core Team. (2014). *R: A language and environment for statistical computing*. Accessed 5 July 2017 at http://www.r-project.org/

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

Revelle, W. (2016). *Psych: Procedures for psychological, psychometric, and personality research*. Accessed 5 July 2017 at http://personality-project.org/r/book/

Saito, K., Webb, S., Trofimovich, P., & Isaacs, T. (2016a). Lexical correlates of comprehensibility versus accentedness in second language speech. *Bilingualism: Language and Cognition*, *19*, 597–609.

Saito, K., Webb, S., Trofimovich, P., & Isaacs, T. (2016b). Lexical profiles of comprehensible second language speech. *Studies in Second Language Acquisition*, *38*, 677–701.

Salsbury, T., Crossley, S. A., & McNamara, D. S. (2011). Psycholinguistic word information in second language oral discourse. *Second Language Research*, *27*, 343–360.

Sawaki, Y. (2013). Factor analysis. In C. A.Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 1–5). West Sussex, UK: Blackwell Publishing.

Schmitt, N. (2014). Size and depth of vocabulary knowledge: What the research shows. *Language Learning*, *64*, 913–951.

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

Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics*, *30*, 510–532.

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

Svartvik, J., & Quirk, R. (1980). *A corpus of English conversation*. Lund, Sweden: Gleerup.

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Harlow, UK: Pearson Education.

Treffers–Daller, J., Parslow, P., & Williams, S. (2016). Back to basics: How measures of lexical diversity can help discriminate between CEFR levels. *Applied Linguistics*, *39*, 302–327.

Witten, I. H., & Frank, E. (2005). *Data mining practical machine learning tools and techniques*. Boston: Morgan Kaufman.

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic Bulletin & Review*, *15*, 971–979.

Zygmont, C., & Smith, M. R. (2014). Robust factor analysis in the presence of normality violations, missing data, and outliers: Empirical questions and possible solutions. *Tutorials in Quantitative Methods for Psychology*, *10*, 40–55.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.