Studies in Second Language Acquisition (2025), 1–25
doi:10.1017/S0272263125000075

RESEARCH ARTICLE

The impact of collocational proficiency features on expert ratings of L2 English learners' writing

Ben Naismith[1,2] and Alan Juffs[1]

[1]Department of Linguistics, University of Pittsburgh, Pittsburgh, PA, USA and [2]Duolingo, Pittsburgh, PA, USA
Corresponding author: Ben Naismith; Email: ben.naismith@duolingo.com

(Received 17 December 2023; Revised 29 November 2024; Accepted 07 January 2025)

Abstract

Lexical proficiency is a multifaceted phenomenon that greatly impacts human judgments of writing quality. However, the importance of collocations' contribution to proficiency assessment has received less attention than that of single words, despite collocations' essential role in language production. This study, therefore, investigated how aspects of collocational proficiency affect the ratings that examiners give to English learner essays. To do so, collocational features related to sophistication and accuracy were manipulated in a set of argumentative essays. Examiners then rated the texts and provided rationales for their choices. The findings revealed that the use of lower-frequency words significantly and positively impacted the experts' ratings. When used as part of collocations, such words then provided a small yet significant additional boost to ratings. Notably, there was no significant effect for increased collocational accuracy. These findings suggest that low-frequency words within collocations are particularly salient to examiners and deserving of pedagogic focus.

Keywords: accuracy; collocation; collocational proficiency; sophistication; writing assessment

Essay writing remains a central component of high-stakes English language proficiency (ELP) exams, which determine whether test takers can study or work in English-medium contexts. It is therefore crucial to understand the specific linguistic features that expert raters attend to when evaluating these essays. One important indicator of lexical proficiency is the use of formulaic sequences, that is, any string of words that can be identified or usefully thought of as a single lexical unit (Siyanova-Chanturia & Pellicer-Sánchez, 2020). Examples include multiword verbs (e.g., *come up with*), idioms (e.g., *under the weather*), and collocations (e.g., *sheepish grin*). Such sequences are potentially even more important than single words in predicting text quality (Bestgen, 2017) and account for a substantial portion of expert-level production (Conklin & Schmitt, 2012; Siyanova-Chanturia & Martinez, 2015).

Of the many types of formulaic sequences, collocations possess a unique status in frameworks of lexical knowledge (e.g., Nation, 2013; Read, 2004), and their importance

in writing proficiency is widely acknowledged (e.g., Crossley et al., 2015; Durrant, 2019). However, few studies have specifically and systematically examined the impact of different collocational features on expert raters' judgments of essays. This study therefore investigated this topic, focusing on dimensions of collocational proficiency, specifically one aspect of collocational sophistication (collocate frequency) and collocational accuracy. Based on these data, the paper makes recommendations for incremental changes to curricula and rating rubrics.

Defining and identifying collocations

Simplistically, collocations are word partnerships that may consist of various linguistic patterns (Szudarski, 2023) such as verb+article+noun (e.g., *break the spell*) or adverb +adjective (e.g., *utterly ridiculous*). As such, they occupy an interesting intermediate space between lexis and syntax (Nattinger & DeCarrico, 1992), possessing both concrete (vocabulary-like) and abstract (grammar-like) qualities.

However, pinpointing exactly what constitutes a collocation depends on the approach taken; the two most common are *the phraseological approach* and *the frequency-based approach*. The phraseological approach uses syntactic, semantic, and pragmatic linguistic criteria (Henriksen, 2013; Lundell & Lindqvist, 2012). For example, a distinction might be whether or not a word combination is more compositional in nature, as in *pay the bill*, or more figurative, as in *pay attention* (Wolter, 2020). In contrast, the frequency-based approach treats the probability of co-occurrence of words as of paramount importance (Henriksen, 2013). Such co-occurrence is often measured by Mutual Information (MI) and t-score, though numerous other measures also exist (e.g., Delta P and Log Dice). A typical convention is to consider word combinations with an MI score over 3 or a t-score over 2 to be a collocation (Church & Hanks, 1990; Jiang, 2009), in conjunction with a minimum frequency threshold from 5 to 10 occurrences of the word combination in the corpus (e.g., Granger & Bestgen, 2014; Simpson-Vlach & Ellis, 2010).

It is possible to combine phraseological and frequency-based approaches by starting with computational extraction for frequency measures and then subsequently applying phraseological criteria (e.g., Laufer & Waldman, 2011; Naismith & Juffs, 2021). Adopting a combined approach admits two key elements of collocations: (1) the frequency with which the words occur together, and (2) the semantic link between the words. Both of these elements can be seen in the definition by Laufer and Waldman (2011, p. 648) which we adopt in the current study:

> [Collocations are] habitually occurring lexical combinations that are characterized by restricted co-occurrence of elements and relative transparency of meaning.

In this conceptualization of collocation, "habitually occurring" combinations can be measured statistically with a frequency-based approach and "restricted co-occurrence and relative transparency of meaning" with a phraseological perspective.

Lexical proficiency and collocations

In its broadest sense, lexical proficiency is "an ability to apply both declarative and procedural lexical knowledge in real language use" (Lenko-Szymanska, 2019, p. 39).

https://doi.org/10.1017/S0272263125000075 Published online by Cambridge University Press

With respect to knowledge and use of collocations (i.e., collocational proficiency), research has shown that expert speakers and learners differ substantially (Granger & Bestgen, 2014; Siyanova-Chanturia & Sidtis, 2019). Learners overuse collocations that they know well (Granger, 1998; Laufer & Waldman, 2011), but underuse collocations more generally, both in quantity and range (Durrant & Schmitt, 2009; Tsai, 2015). The reason for such failures is likely a combination of factors that may include collocations' relative infrequency in input (Gyllstad & Wolter, 2016), the lack of a literal counterpart in the learner's L1 (Macis & Schmitt, 2016), their lack of salience as linguistic items (Lee, 2019; Wolter, 2020), or how they were taught (Jiang, 2009; Siyanova-Chanturia & Spina, 2020). Collocational proficiency is, therefore, one factor that can distinguish among levels of proficiency (Ha, 2013; Lundell & Lindqvist, 2012).

Two important dimensions of lexical and collocational proficiency that have been shown to impact perceptions of text quality are sophistication and accuracy. While both dimensions are frequently considered in relation to the use of single words, they also apply to the use of formulaic sequences such as collocations. We now discuss sophistication and accuracy in relation to single words and collocations.

Lexical and collocational sophistication

Lexical sophistication commonly refers to the use of advanced or sophisticated words (Kim et al., 2018) that reflect the breadth and depth of lexical knowledge (Kyle & Crossley, 2015). By nature, lexical sophistication is multidimensional. For example, in Eguchi & Kyle's (2020) framework, the construct includes rareness (frequency and dispersion), conceptual features (e.g., concreteness), distinctiveness, accessibility, and association measures of multiword units. Of these, rareness remains the most commonly investigated through the use of frequency-based measures related to the proportion of relatively advanced words produced in a text (Read, 2000). Here, we restrict our focus to frequency measures due to their relevance to the current study and usefulness for simultaneously considering both single words and collocations.

Numerous studies have found that indices of lexical sophistication correlate with human judgments of writing (e.g., Eguchi & Kyle, 2020; Kim et al., 2018; Lenko-Szymanska, 2019; Vögelin et al., 2019). For example, Lenko-Szymanska (2019) found that three frequency-based measures of lexical sophistication—the percentage of words beyond the 2000 most frequent words, the percentage of academic words, and the mean log frequency of content words—were able to discriminate well between texts written by learners of different proficiency levels. Vögelin et al. (2019) likewise found a positive relationship between human ratings and scores with higher frequency-based sophistication. Manipulating the lexical sophistication of texts, as measured by average word range, these researchers found that texts with greater lexical sophistication received significantly higher scores from teachers for vocabulary ($\eta^2 = .348$, $p < .001$) as well as for holistic quality ($\eta^2 = .110$, $p < .05$). Exploring the impact of both single-word and multiword lexical indices, Kim et al. (2018) found that models of lexical sophistication combining both types of frequency indices were the most predictive of scores of L2 writing (24.6% of variance) and lexical proficiency (31% of variance). Single-word indices were related to the use of advanced words (including frequency, dispersion, and psycholinguistic properties), and multiword indices were related to association measures and frequency measures.

To discuss frequency-based sophistication measures in relation to collocations, it is necessary to first differentiate between two types of collocational frequency. The first

https://doi.org/10.1017/S0272263125000075 Published online by Cambridge University Press

can be coined a *collocation frequency approach*, so that, for example, in The Corpus of Contemporary American English (COCA; Davies, 2008), the lemma combination of *needless* and *say* (as in *needless to say*) occurs 3,735 times and has an MI of 4.37. As such, it has a lemma combination rank of 251 and can be considered a high-frequency collocation in comparison to other collocations. Alternatively, we can look at the lemmas individually, that is, through a *collocate frequency approach*, in which case *say* is certainly high frequency (lemma frequency = 4,096,416, lemma rank = 26), but *needless* is much lower frequency (lemma frequency = 4,942, lemma rank = 8,468). Thus, if a learner uses *needless to say* in an essay, should this collocation be considered evidence of low sophistication because it is a common collocation or high sophistication because it is a collocation containing an uncommon word?

These two views of collocational sophistication reflect the nature of collocations as simultaneously holistic chunks and also compositional strings of words. This dual view is supported by processing studies which have shown that formulaic sequences become increasingly prominent as single units through repeated use, but that they retain information about their parts (Öksüz et al., 2021; Wolter & Yamashita, 2018). Of the two, the collocation frequency approach is more common and perhaps more intuitive. And yet, findings regarding the relative importance of collocational frequency have been mixed. In a meta-analysis of 19 collocation studies, Durrant (2014) found that collocation frequency correlated only moderately with collocation knowledge; other important factors included semantic transparency and the amount of social engagement of learners. However, in studies by Garner and colleagues (Garner 2022; Garner et al., 2019, 2020), the more proficient writers were observed to use lower-frequency collocations, for example, more sophisticated verb-noun collocations (Garner, 2022).

Studies using a collocate frequency approach have been more interested in collocations in relation to the individual collocates contained within them. For example, Ebrahimi (2017) investigated the collocational knowledge of Iranian EAP learners, specifically collocations composed of high-frequency words. Jiang (2009) focused on pedagogic materials for teaching collocations to Chinese learners and found that 93.6% of collocates belonged to the K1-2 frequency bands. González Fernández and Schmitt (2015) incorporated both approaches and looked at the link between frequency and productive collocation knowledge, but only for collocations whose collocates were in the K1-5 frequency bands. Matching Durrant (2014), the study found only a weak relationship between collocation frequency and collocation knowledge. Several other studies reporting the association measure of MI have also demonstrated that higher MI correlates with higher learner proficiency (e.g., Granger & Bestgen, 2014; Jiang et al., 2023; Paquot, 2018). Although the focus on MI in these works has been to investigate the degree of association in word pairs, note that word (or lemma) frequency is also part of the MI equation. As a result, low-frequency words often result in more exclusive combinations and consequently receive higher MI scores (Szudarski, 2023), indicating a relationship between MI, collocate frequency, and learner proficiency.

This paper focused on using the collocate frequency approach to be able to classify collocations as *low-*, *mid-*, or *high-frequency* based on single-word frequency statistics from external corpora. In doing so, we were better able to compare the effects of frequency on text quality in relation to both single words and collocations containing those words. To our knowledge, no studies have yet to apply the *mid-frequency* label to collocations, and the labels *high-frequency* and *low-frequency* have been used variably (e.g., Durrant & Schmitt, 2009; Yoon, 2016).

While word frequency is an interval variable, as evidenced in the studies above, it is commonly partitioned into frequency bands of 1,000 (K) words, or "K-bands." Many

https://doi.org/10.1017/S0272263125000075 Published online by Cambridge University Press

authors have also suggested a three-way distinction between high-, mid-, and low-frequency lexical items (e.g., Naismith & Juffs, 2021; Vilkaitė-Lozdienė & Schmitt, 2020). In this format, common practice based on coverage statistics defines the K1-2 frequency bands as high-frequency, K3-9 as mid-frequency, and K10+ as low-frequency. There have been calls for other bucket sizes, for example, Kremmel's (2016) suggestion of 500-item bands for K1-3, 1,000-item bands for K4-6, and 2,000-item bands for K7-10. It is true that operationalizing frequency as 1,000-item bands can lose more fine-grained information, but there are pedagogical and research advantages to establishing such categories. For example, for learners wishing to study in an L2 academic environment, identifying and learning mid-frequency lexis is particularly important (Nation & Anthony, 2013; Vilkaitė-Lozdienė & Schmitt, 2020) as it is essential for achieving sufficient coverage of academic texts (Laufer, 1989; Laufer & Ravenhorst-Kalovski, 2010). Knowing what lexis is mid-frequency, therefore, allows for clear learning goals to be set which are easily accessible to learners, teachers, and materials developers.

Lexical and collocational accuracy

Simply put, lexical accuracy is the ability to produce writing free from lexical errors. In general, there is a strong negative correlation between the number of errors and holistic ratings (Polio & Shea, 2014), and lexical errors have been found to occur more than grammatical errors (Agustín Llach, 2011; Qian & Lin, 2020). Because lexical errors affect communication, they are highly prominent and are therefore judged more severely by readers and listeners (Ellis, 2008; Santos, 1988). The typical quantitative approach to accuracy is to count either error-free units, like T-units or clauses (e.g., Polio, 1997), or errors themselves (e.g., Linnarud, 1986). These counts can then be normalized using ratios such as the number of errors per word, per lexical word (typically nouns, adjectives, verbs, adverbs), or per 100 words.

With respect to collocations, "free from lexical errors" can refer to whether these combinations are acceptable and expected (Crossley et al., 2013). Collocational accuracy is especially important in academic writing as collocation misuse indicates a lack of academic expertise (Henriksen, 2013) and forces readers to decompose the collocations rather than process them fluently as single chunks (Howarth, 1998). Even if the meaning of the individual words is not obscured by how they are combined, collocation errors can still strain the reader through "lexical dissonance" (Hasselgren, 1994), increasing the processing burden (Millar, 2011).

Numerous studies have demonstrated the high prevalence of collocational errors at all proficiency levels of L2 English writing. For example, approximately 33% of the collocations investigated by Laufer and Waldman (2011) and 50% by Nesselhauf (2005) were incorrect. There is also a strong case for the impact of collocational accuracy on human judgments of proficiency. In Crossley et al. (2015), collocational accuracy explained 84% of the variance in human judgments between the writing samples and was one of the three most predictive variables. In addition, the studies showing a positive relationship between collocation association measures (like MI) and proficiency can be considered indirect evidence of the importance of collocational accuracy since lower MI may be indicative of higher rates of inappropriate word choice. It should be noted, however, that in Laufer and Waldman (2011), similar rates of collocational errors were seen at all proficiency levels. Other factors that may affect collocation accuracy rates include the definition of potential collocations, the types of collocations under investigation, and the L1s of the learners.

https://doi.org/10.1017/S0272263125000075 Published online by Cambridge University Press

Human rating of writing

Thus far, we have discussed how statistical measures of single- and multiword lexical items correspond to language proficiency, without consideration of how language proficiency is measured. Commonly, assessments of writing proficiency rely on human raters' variable and subjective perceptions of quality, characteristics that impact validity and reliability (Attali, 2016; Eckes, 2012). Despite these human factors, such assessment is still widely administered because it directly tests communicative language ability (Hamp-Lyons, 1990), in an approach where some "errors" are tolerated as part of a global approach to success in language ability and not penalized as in earlier thinking on testing.

The reasons for rater variability are legion. Assessing essays imposes a high cognitive demand (Eckes, 2012), and even looking at lexis alone, assigning a numeric score is a challenge (Fritz & Ruegg, 2013). Ratings can vary across raters (inter-rater reliability) or may "drift" for one rater across texts (intra-rater reliability). Raters perhaps differ in severity/leniency because their perceptions of the importance of various criteria vary (Eckes, 2012; Goh & Ang-Aw, 2018; Lumley & McNamara, 1995). In addition, research on any potential advantage for raters' experience is less clear (e.g., Cumming, 1990; Lim, 2011), though rater training can improve intra-rater reliability and adherence to rubrics (Brown, 2006; Hall & Sheyholislami, 2013). Statistical models such as Many-Facet Rasch Measurement models (MFRM; Linacre, 1989, 1994) can also be used to account for systematic rater severity/leniency (see McNamara et al., 2019).

Particularly relevant to this paper, think-aloud studies and retrospective comments after grading suggest that lexis has not been of primary consideration for some raters (Goh & Ang-Aw, 2018; Lumley & McNamara, 1995), even though evaluation rubrics may include a vocabulary category. However, using think-aloud protocols alters the thought process of the raters (Barkaoui, 2011; Lumley, 2005), so any conclusions in this regard must be considered tentative. Raters may also perceive longer texts to be of superior quality and give them higher ratings just for that reason alone (Guo et al., 2013; Kyle et al., 2020; Linnarud, 1986). Therefore, text length can "wash out" the predictive strength of other lexical variables (Crossley & McNamara, 2012) and should be controlled for.

Still, the relationship between lexical features and human judgments has long been a focus of writing assessment research, as exemplified by studies discussed in the previous section (Kim et al., 2018; Vögelin et al., 2019). Early investigations of lexical measures in L2 writing (e.g., Arnaud, 1984; Linnarud, 1986) established that features like lexical diversity, sophistication, and accuracy can distinguish proficiency levels. More recent research has investigated specific dimensions of lexical proficiency and their impact on assessments. For instance, Bestgen and Granger (2014) found that essays with more sophisticated collocations (measured by MI scores) received higher ratings; Leńko-Szymańska (2019) showed that raters attend to different aspects of lexical proficiency in their evaluations; Lu and Hu (2022) demonstrated that sense-aware lexical sophistication indices improved prediction of writing quality over traditional indices; and Monteiro et al. (2020) found that L2 lexical sophistication indices were significantly stronger predictors of holistic ratings than L1 benchmarks, explaining twice the variance. Studies like these highlight the importance of lexical features in human judgments and the many ways in which dimensions of lexical proficiency can be operationalized. Additionally, recent meta-analyses have examined the relationships between L2 writing performance and its internal and external correlates. Of relevance here, moderate correlations were found between L2 writing performance and lexical

https://doi.org/10.1017/S0272263125000075 Published online by Cambridge University Press

complexity (r = .295; Kojima & Kaneta, 2022) and L2 vocabulary knowledge (r = .489; Kojima et al., 2022). These meta-analytic findings reinforce the contribution of lexical features to assessments of L2 proficiency.

Two previous human rating studies are especially noteworthy to the context of this paper. First, Fritz and Ruegg (2013) focused on argumentative essays written under timed conditions. However, rather than analyzing a wide range of essays, a single "base" essay was used. The 32 content words in this base text were manipulated to create 27 total versions: low/mid/high versions of accuracy, diversity, and sophistication, that is, a 3×3 design. Twenty-seven experienced raters used four analytic scales to assess the essays, and these ratings were analyzed using analyses of variance (ANOVAs) to find the relationships between variables. The findings indicated that lexical accuracy significantly predicted ratings, $F(2, 68) = 4.262$, $p = .013$, though surprisingly diversity, $F(2, 68) = .69$, $p = .933$, and sophistication, $F(2, 68) = 1.68$, $p = .194$, did not. Importantly, the authors acknowledged certain limitations: experimental texts were mixed with "authentic" texts (which affected the ratings) and the operationalization of sophistication was somewhat problematic. From that study, this paper adopted several approaches to experimental control.

Second, Read and Nation (2006) investigated the vocabulary use of International English Language Testing System (IELTS) test takers. Their study analyzed 88 recordings of learners completing their Part 2 "long turns." Speech with higher ratings contained a higher percentage of low-frequency vocabulary, and qualitatively, at the highest levels, was characterized by "mastery of colloquial or idiomatic expressions." This result points to the importance of both single- and multiword lexical items to examiners, as well as the need for further research of IELTS lexical resource ratings.

Current study

The goal of the current study is to isolate and measure the contributions of collocational features to overall ratings of lexical resource quality in essays by comparing quantitative text metrics, expert ratings, and the rationales for these ratings. Three research questions are addressed:

1. To what extent are expert ratings of lexical proficiency of essays impacted by
   a. the number of high-/mid-/low-frequency lemmas (a dimension of lexical sophistication)?
   b. whether or not the high-/mid-/low-frequency lemmas are part of collocations (a dimension of collocational proficiency)?
2. To what extent are expert ratings of lexical proficiency of essays impacted by the number of accurate and inaccurate collocations (a dimension of collocational accuracy)?
3. What aspects of lexical proficiency do the expert raters consciously attend to, as reflected in their comments, and do these include aspects of collocational proficiency?

Investigating these questions is significant for enhancing our understanding of the linguistic features that expert raters attend to when evaluating L2 writing. As we have seen, one line of previous research has demonstrated the multidimensional nature of lexical sophistication and its impact on judgments of proficiency (e.g., Kim et al., 2018; Lenko-Szymanska, 2019). Other lines of inquiry have shown how aspects of

https://doi.org/10.1017/S0272263125000075 Published online by Cambridge University Press

collocational sophistication, such as the use of collocations with higher MI, correlate with higher learner proficiency (e.g. Granger & Bestgen, 2014; Paquot, 2018). Understanding these relationships is particularly relevant for writing assessment, where current descriptors, including those used by IELTS, vary in how explicitly they address different aspects of collocational proficiency.

The present study aimed to extend the lines of inquiry above by examining how variations in single-word and collocational features impact expert ratings of lexical proficiency, using a carefully controlled dataset of written texts designed to systematically manipulate these features. Given the centrality of collocation knowledge in frameworks of lexical proficiency (e.g., Nation, 2013), it is crucial to determine the extent to which collocational features impact raters' judgments and to better understand how the impact of lemma frequency on expert judgments is mediated by placement within collocations. Furthermore, by comparing raters' qualitative comments with frequency-based measures of lexical/collocational sophistication and collocational accuracy, this study can provide insight into the alignment between the features that raters consciously notice and those that statistically predict their scores. The findings thus have implications for identifying areas of convergence and divergence between theoretical constructs of collocational proficiency and raters' actual practices, providing a more nuanced understanding of the linguistic features that shape expert judgments of L2 writing quality.

Methods

This study used an embedded design in which both quantitative data (the ratings) and qualitative data (the reflections) were collected simultaneously (Creswell & Plano Clark, 2011), with the reflections enhancing the completeness of the data. First, raters accessed a link for viewing/downloading the rating scales and task prompts. Next, they rated three texts at different Common European Framework for Reference of Languages (CEFR) levels. After rating each text, the raters answered follow-up questions about their assessment decisions. Finally, raters provided personal metadata. To ensure the validity of these findings, a large number of expert raters were used; the texts rated were identical in length and topic; and rater effects were controlled for through the use of MFRM models.

Participants (raters)

Because the target population is raters who evaluate high-stakes tests, participation in the study was limited to current or former IELTS examiners. All IELTS examiners must meet minimum requirements of substantial (typically 3+ years) teaching experience to adults, an undergraduate degree, a recognized TEFL/TESOL qualification or degree in education, and expert spoken and written English proficiency. IELTS examiners undergo a comprehensive training and certification process, as well as subsequent monitoring and standardization. To recruit the raters, snowball sampling was used, a type of sampling of convenience that is a well-established practical option for recruiting members from hard-to-reach groups (Valdez & Kaplan, 1998).

To determine the required number of participants, an *a priori* power analysis was performed using G*power (version 3.1.9.6; Faul et al., 2007). For linear multiple regression, to detect a small effect size (d = .2; Cohen, 1988) with a power of .8 and α of .05, 40 raters were required given the experimental design. In total, there were

https://doi.org/10.1017/S0272263125000075 Published online by Cambridge University Press

Table 1. Rater information (IELTS examiners)

n = 47						
Gender	*Man*	*Woman*	*Unknown*			
	24	20	3			
Age	*30–39*	*40–49*	*50–59*	*60–69*	*≥ 70*	
	10	14	15	7	1	
Education	*BA*	*MA*	*PhD*			
	7	36	4			
TESOL certification	*Certificate*	*Diploma*	*Degree*	*Other*		
	11	31	4	1		
TESOL experience (years)	*6–10*	*11–20*	*> 20*			
	3	17	27			
IELTS status	*Examiner*	*Examiner trainer*				
	38	9				
IELTS experience (years)	*< 1*	*1–2*	*3–5*	*6–10*	*11–20*	*> 20*
	1	6	13	7	15	5
Rater L1	*English*	*Other*				
	38	9				
Student proficiency range	*Wide*	*Narrow*				
	33	14				
Student L1 range	28	19				

48 respondents, though one was excluded as they appeared to be a non-examiner based on their responses. This participant pool size had the desired effect of allowing multiple raters for each script.

Table 1 presents the raters' demographic information. These data represent mature examiners in terms of age (all over 30) and experience, both TESOL experience (94% have > 10 years' experience) and examining experience (85% have > 2 years' experience). Of the 47 participants, 19% were examiner trainers, held to a higher standard of reliability. Most commonly, the participants had experience with a wide range of first languages (60%) and proficiency levels (70%). The participants were also highly educated, with most possessing graduate degrees (85%) and additional TESOL certification (89%).

Instruments

Three initial texts formed the basis for the texts in the survey. These are IELTS Task 2 responses (IELTS, n.d.-a), selected because they are publicly available and accompanied by examiner ratings and comments. All three texts responded to the same task about the relationship between socioeconomic status and problem-solving ability. The overall scores of the three texts are Bands 4, 6.5, and 8, displaying a wide range of proficiency levels on the IELTS scale of 1 to 9. These scores correspond to CEFR levels of B1, B2, and C1, respectively. Although originally handwritten, the texts were typed for practicality and standardization purposes, and only two orthographic errors were corrected in the B1 text. There was no background information on the writers.

To control for the issue of text length, the three original texts were normalized to 250 words through careful manual alterations, endeavoring to maintain all stylistic aspects of the original texts. Throughout the process of text manipulation, precise quantitative analysis of the texts was carried out to ensure that 15 key lexical, syntactic, and collocational metrics remained within 5% of the original texts. Collocations were identified using a combined phraseological and frequency-based approach: a checklist

https://doi.org/10.1017/S0272263125000075 Published online by Cambridge University Press

Table 2. Collocational density of text versions

Text	Length	Accurate cols	Accurate cols per 100	Inaccurate cols	Inaccurate cols per 100
B1 initial	172	8	4.7	14	8.1
B1 normalized	250	12	4.8	20	8.0
B1 final	**250**	**12**	**4.8**	**18**	**7.2**
B2 initial	349	31	8.9	11	3.2
B2 normalized	250	22	8.8	8	3.2
B2 final	**250**	**22**	**8.8**	**12**	**4.8**
C1 initial	254	33	13.0	5	2.0
C1 normalized	250	33	13.2	5	2.0
C1 final	**250**	**32**	**12.8**	**6**	**2.4**

of phraseological criteria was first used by both authors.[1] Criteria based on frequency statistics from the COCA corpus (Davies, 2008) were then used to settle disagreements and to filter potential collocations ($n > 5$, MI > 3, t-score > 2). Each collocation occurred once per text. Likewise, inaccurate collocations were first identified by authors as word combinations where a collocation was expected, but the word combination did not meet the checklist criteria, standing out as an unnatural/awkward/unclear choice. These inaccurate collocations were then confirmed to not meet the frequency statistics listed above.

Once normalized, a subsequent step of text manipulation was carried out to more evenly space accurate and inaccurate collocational use across the texts (Table 2). To ensure that these manipulations did not affect the initial IELTS scores/CEFR levels, a pilot study was carried out to rate the initial, normalized, and final texts. Overall scores were calculated as an average of the four analytic bands and rounded down to the nearest .5, following IELTS practices. These scores indicate that at all three proficiency levels, the normalization and manipulation processes did not greatly impact the average ratings, with all changes within half a band, maintaining the original CEFR levels (Figure 1). An analysis of the analytic bands revealed similar patterns.

Using the three final texts, 30 different versions were created (10 versions per final text) by changing up to approximately 12% of the words.[2] These manipulations were intended to influence several collocational indices relating to sophistication (Bestgen & Granger, 2014; Granger & Bestgen, 2014) and accuracy:

1. *Mean MI*: to measure association of collocations containing infrequent words (formula from Davies, 2008)
2. *Mean t-score*: to measure association of collocations containing high-frequency words (formula from Evert, 2009)
3. *Absent bigrams*: the proportion of bigrams absent from the reference corpus
4. *Accurate collocations and collocation errors:* number per 100 words, based on error types in Granger (2003) and Wanner et al. (2013). Collocations present in the task prompt were not counted.
5. *Collocation frequency bands*: to determine whether each collocation contained only high-frequency lemmas (K1-2), a mid-frequency lemma (K3-9), or a low-frequency lemma (K10-16). Frequency bands were determined by ranking COCA lemma

[1]Available online as supplementary materials.
[2]Available online as supplementary materials.

https://doi.org/10.1017/S0272263125000075 Published online by Cambridge University Press

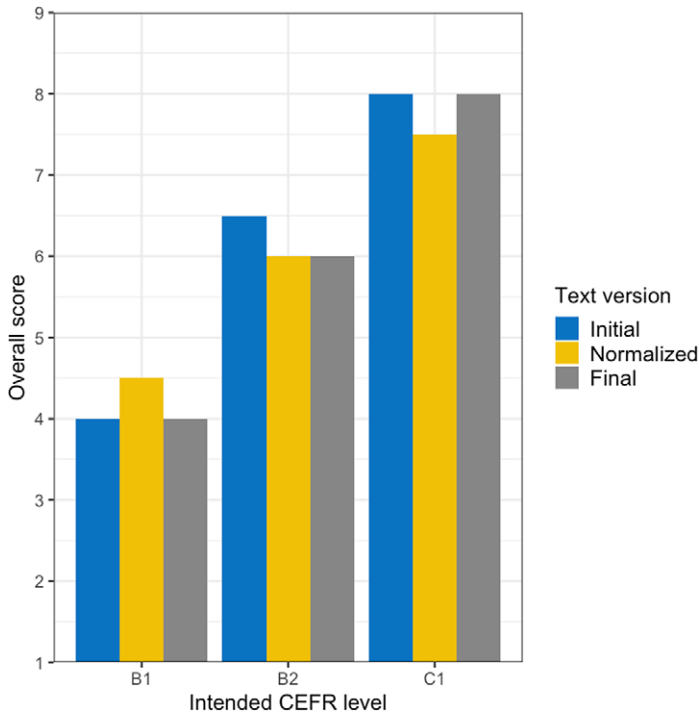

Figure 1. Overall ratings comparison of initial, normalized, and final texts.

frequencies (available at https://www.wordfrequency.info/purchase.asp). The proportion of each of these types of collocations in the text was then calculated.

The overarching selection criteria for the indices was the "meaningfulness and interpretability of the information they encapsulate as well as their theoretical motivation" Lenko-Szymanska, 2019, p. 161). These indices correlate with human judgements of proficiency and align with the lexical subconstructs evidenced in the IELTS Task 2 band descriptors (IELTS, n.d.-b). As a result of the text manipulations, the text versions differed in terms of four variables: proficiency level, collocational (collocate) frequency, non-collocational lemma frequency, and collocational accuracy. Table 3 presents a matrix of all 30 text versions.

1. *Proficiency level:* Three CEFR proficiency levels, B1 (intermediate), B2 (upper-intermediate), C1 (advanced).
2. *Collocational (collocate) frequency:* Three levels, High, Mid, and Low frequency. To change the levels, accurate collocations were replaced based on the lemma frequencies of the collocates from COCA (Davies, 2008) and verified as "basic" or "advanced" lemmas in the PELIC learner corpus (Juffs et al., 2020; Naismith et al., 2022), for example, high = *good example* (K1) → mid = *concrete example* (K4).

https://doi.org/10.1017/S0272263125000075 Published online by Cambridge University Press

Table 3. Characteristics of text versions

Text	Proficiency	Frequency in COCA	Accuracy	Accurate cols				Col errors
				K1-2	K3-9	K10-16	Total	
1	B1	High	Low	9	3	0	12	18
2	B1	High	High	14	4	0	18	12
3	B1	Mid (collocational)	Low	0	12	0	12	18
4	B1	Mid (non-collocational)	High	5	13	0	18	12
5	B1	Mid (collocational)	Low	9	3	0	12	18
6	B1	Mid (non-collocational)	High	14	4	0	18	12
7	B1	Low (collocate)	Low	0	0	12	12	18
8	B1	Low (collocate	High	5	1	12	18	12
9	B1	Low (non-collocational)	Low	9	3	0	12	18
10	B1	Low (non-collocational)	High	14	4	0	18	12
11	B2	High	Low	16	6	0	22	12
12	B2	High	High	20	8	0	28	6
13	B2	Mid (collocational)	Low	4	18	0	22	12
14	B2	Mid (non-collocational)	High	8	20	0	28	6
15	B2	Mid (collocational)	Low	16	6	0	22	12
16	B2	Mid (non-collocational)	High	20	8	0	28	6
17	B2	Low (collocate)	Low	7	3	12	22	12
18	B2	Low (collocate	High	11	5	12	28	6
19	B2	Low (non-collocational)	Low	16	6	0	22	12
20	B2	Low (non-collocational)	High	20	8	0	28	6
21	C1	High	Low	18	11	3	32	6
22	C1	High	High	22	13	3	38	0
23	C1	Mid (collocational)	Low	8	23	1	32	6
24	C1	Mid (non-collocational)	High	12	25	1	38	0
25	C1	Mid (collocational)	Low	18	11	3	32	6
26	C1	Mid (non-collocational)	High	22	13	3	38	0
27	C1	Low (collocate)	Low	11	6	15	32	6
28	C1	Low (collocate	High	15	8	15	38	0
29	C1	Low (non-collocational)	Low	18	11	3	32	6
30	C1	Low (non-collocational)	High	22	13	3	38	0

3. *Non-collocational frequency:* The same characteristics of collocational frequency apply to non-collocational frequency. The only difference was that the words altered were not part of collocations, for example, mid = *nevertheless* (K4) → low = *unbelievably* (K14).
4. *Collocational accuracy:* Two accuracy levels, *Low* and *High.* At each proficiency level, there were six additional inaccurate collocations in the low level. For example, Text 1 is a B1 level, has 12 accurate collocations from two frequency bands, and has 18 collocations with errors.

To assess the texts analytically, raters used the IELTS public writing scales (IELTS, n.d.-b). For each of the four categories—Task Response (TR), Coherence and Cohesion (CC), Lexical Resource (LR), and Grammatical Range and Accuracy (GRA)—there were bands from 1 to 9 with descriptive criteria. For this study, the 9-point band scale was further divided into three sublevels (e.g., 5-, 5, 5+) so that there was a "strong" and "weak" possibility within each band (a practice used in Jarvis, 2013) to increase the range of possible ratings. For data analysis, these sublevels were converted to decimals, so that, for example, 5- → 5.0, 5 → 5.3, and 5+ → 5.7. In addition to the analytic scales, the raters provided a holistic assessment based on the IELTS public 9-band overall scale

https://doi.org/10.1017/S0272263125000075 Published online by Cambridge University Press

(IELTS, n.d.-c). IELTS examiners do not give holistic assessments, but this additional holistic rating served to align the methods of the current study with other comparable research and provided an extra level of data for analysis. The holistic scales were minimally adapted to remove reference to spoken production and language comprehension.

Results

In this section we present the analysis of the ratings data (Research questions 1 and 2) and survey data (Research question 3) to determine how aspects of lexical and collocational proficiency impacted the ratings of lexical proficiency.

Quantitative data analysis

We first used an MFRM model to arrive at *fair scores* for each text, that is, the rating that would have been given by a rater of average severity. In doing so, we sought to mitigate the systematic error inherent in human ratings. In essence, MFRM models "predict the outcome of encounters between persons and assessment/survey items" (Aryadoust et al., 2021, p. 7) by considering multiple variables (referred to as *facets*). Here, a three-facet model was created using FACETS software (Linacre, 2020), consisting of the raters, the texts, and the band descriptors. In addition, other distal factors (demographic and task variables) were tested, but none of these variables indicated significant bias. The output of the model was the ratings expressed in log-odds units (logits), which were then transformed into the fair scores.[3]

Having established fair scores, the impact of collocational features on the lexical ratings could be calculated using a linear regression model created in the R environment (version 3.6.2; R Development Core Team, 2019). Prior to creating the model, the assumptions required by linear regressions were checked and met (Levshina, 2015). In the model (Table 4), the outcome variable is the Lexical Resource fair scores (LR_fair). The independent variables are the fair scores for the other analytic criteria (TR, CC, GRA), the frequency in COCA of manipulated lexical items (High, Mid, Low), the type of manipulated lexical item (Collocation, Non-collocation), the collocation accuracy (Low, High), and the base text CEFR level (B1, B2, C1). In addition, motivated interactions were included. In this experimental design, all potential variables were left in the model regardless of whether they improved the model fit. The independent variables were sum contrast coded with the exception of frequency; the reference level for frequency is therefore High. As a result of the contrast coding, the model's intercept is dispersed across levels of the other variables. By comparing all categories against the grand mean in this manner, the results are more informative in terms of the deviation of each category from the overall average rather than comparisons to a specific baseline category. However, because such coding focuses on *overall* effect estimation, the model estimates can be difficult to interpret, and it is useful to subsequently use Tukey's Honestly Significant Difference (HSD) test to interpret pairwise comparisons between levels of the variables of interest. For example, for the CEFR variable, in Table 4 we see that the CEFR level (rows 2 and 3) is significant. The post-hoc analysis in Table 5 confirms that the levels are reliably different, increasing as expected from B1 → B2 → C1.

[3]All ratings available online as supplementary materials.

https://doi.org/10.1017/S0272263125000075 Published online by Cambridge University Press

Table 4. Linear regression model for factors predicting lexical resource ratings

Parameters	Estimate	SE	CI	t-value	Pr(>|t|)
(Intercept)	.831	.069	.69–.97	11.975	**< .001*** **
CEFR [B1-C1]	–.180	.018	–.22–-.15	–10.174	**< .001*** **
CEFR [B2-C1]	.070	.007	.06–.08	10.479	**< .001*** **
TR fair	.883	.144	.60–1.17	6.118	**< .001*** **
CC fair	–.827	.262	–1.35–-.31	–3.151	**.002** **
GRA fair	.879	.199	.48–1.27	4.422	**< .001*** **
Freq [low]	.032	.006	.02–.04	5.104	**< .001*** **
Accuracy [low-high]	–.005	.005	–.01–.00	–1.057	.293
Item type [col-non-col]	–.022	.004	–.03–-.01	–4.887	**< .001*** **
CEFR [B1-C1] * freq [low]	–.014	.010	–.03–.01	–1.411	.162
CEFR [B2-C1] * freq [low]	–.010	.009	–.03–.01	–1.182	.240
CEFR [B1-C1] * accuracy [low-high]	–.030	.005	–.04–-.02	–6.454	**< .001*** **
CEFR [B2-C1] * accuracy [low-high]	.001	.005	–.01–.01	.320	.750
Freq [low] * accuracy [low-high]	.009	.006	–.00–.02	1.452	.150
Freq [low] * item type [col-non-col]	.008	.006	–.00–.02	1.367	.175

Note: * $p < .05$, ** $p < .01$, *** $p < .001$.
Model formula: lm(formula = LR_fair ~ CEFR + TR_fair + CC_fair + GRA_fair + freq + accuracy + item_type + CEFR:freq + CEFR: accuracy + freq:accuracy + freq:item_type).

Table 5. Tukey's multiple comparison of means test for CEFR

Contrast	Estimate	SE	CI	df	t ratio	Pr(>|t|)
B1–B2	.252	.017	.22–.29	89	–14.551	**< .001*** **
B1–C1	.310	.031	.25–.37	89	–9.952	**< .001*** **
B2–C1	.058	.017	.00–.06	89	–3.398	**< .003** **

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

RQ1: Collocational sophistication (collocate frequency)

The regression data answered RQ1 which asked the extent to which expert ratings of lexical proficiency of essays are impacted by the number of high-/mid-/low-frequency collocates and non-collocates (a dimension of lexical/collocational sophistication). Overall, there is a significant positive increase, with lower frequency lexis predicting a higher LR rating. The post hoc analysis (Table 6) showed that the bulk of the frequency effect occurred when going from high- to low-frequency. The difference between high- and mid-frequency was also significant ($p = .018$), but there was no significant difference between mid- and low-frequency ($p = .158$).

In addition, there was a significant difference for lexical item type (Table 7), that is, whether lemma frequency effects were mediated by the placement of the lemma within or outside of a collocation. This effect was small but significant, resulting in higher LR ratings when the lower-frequency lemmas were part of a collocation.

Table 6. Tukey's multiple comparison of means test for frequency

Contrast	Estimate	SE	CI	df	t ratio	Pr(>|t|)
mid vs. high	.032	.012	.01–.06	111	2.810	**.018***
low vs. high	.050	.010	.03–.07	111	4.831	**< .001*** **
low vs. mid	.018	.009	.00–.04	111	1.934	.158

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

https://doi.org/10.1017/S0272263125000075 Published online by Cambridge University Press

Table 7. Tukey's multiple comparison of means test for item type

Contrast	Estimate	SE	CI	df	t ratio	Pr(>|t|)
col vs. non-col	.036	.006	.024–.05	89	5.744	**.018***

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

RQ2: Collocational accuracy

The regression data answered RQ2 that asked the extent to which expert ratings of lexical proficiency of essays are impacted by the number of accurate and inaccurate collocations. Of the experimental variables, only collocational accuracy was not significant. Furthermore, interactions between accuracy and the CEFR B2-C1 contrast, and accuracy and low frequency were not significant. One significant interaction was present between accuracy and the CEFR B1-C1 contrast. However, after careful plotting and examination of this interaction, this significant relationship appears to be spurious.

RQ3: Rater rationales

Recall that the design of the study specifically called for quantitative and qualitative insights into raters' scoring practices. Thus, the raters' rationales for their scores answered the third research question which asked which aspects of lexical proficiency expert raters consciously attend to, especially in terms of collocational proficiency. Their comments contained both positive and negative elements, and raters routinely used language directly from the band descriptors. Figure 2 presents a tally of the different lexical features commented on. Only the first occurrence of each term for each rater response was counted, and similar terms were collapsed for clarity, so that, for example, the count for the term "formulaic sequence" includes mentions of "chunk" and "multiword expression." Therefore, 25 for "formulaic sequence" means that 25 rationales (corresponding to a minimum of 9 raters and a maximum of 25 raters) mentioned this construct at least once.

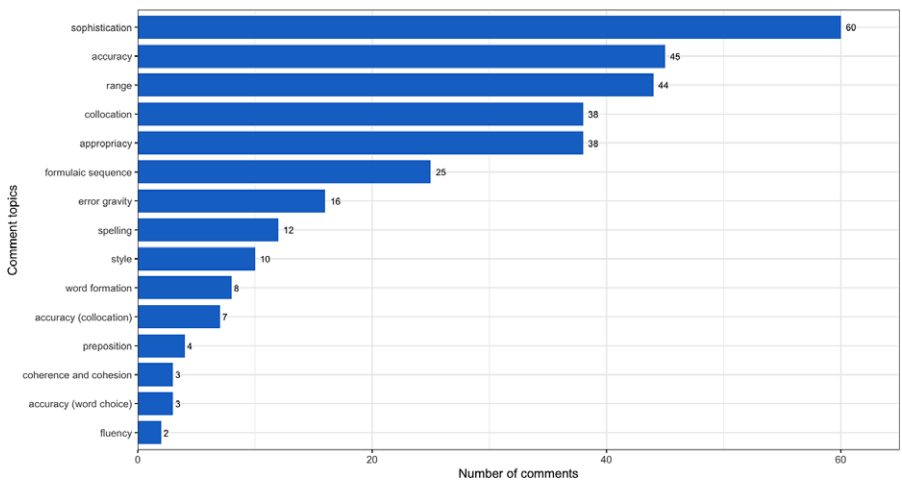

Figure 2. Topics of rater comments.

https://doi.org/10.1017/S0272263125000075 Published online by Cambridge University Press

Here the most common lexical aspects that were noticed correspond to the primary lexical dimensions addressed in this study, sophistication and accuracy, with the importance of collocation also clearly represented. When giving examples, raters tended to give a mix of single and formulaic sequences, thus, the 60 occurrences of the concept of sophistication encompassed both single and multiword lexical units, even though the type of sophistication was not typically described.

Discussion

The findings depict elements of the relationship among different aspects of lexical and collocational proficiency. We first discuss the results in relation to RQs 1 and 2, contextualized with the qualitative results relating to RQ3. We then consider the implications of these results in terms of language pedagogy and assessment.

RQ1: Importance of high-, mid-, and low-frequency collocations

It is unsurprising that the use of lower frequency lemmas (inside and outside of collocations) led to higher ratings of lexical proficiency; this aspect of lexical sophistication, measured in various ways, has long been recognized as a characteristic of more proficient writing (e.g., Daller et al., 2013). The examples given by raters also provided support for the statistical importance of sophistication in terms of frequency; across all three levels, low-frequency single words and collocations containing low-frequency words were flagged as being examples of sophisticated lexis. For example, low-frequency single words repeatedly highlighted by raters in their comments include *fantasize*, *flaunts*, and *tremendous*. Collocations containing mid- or low-frequency collocates included *fairly young*, *first-hand experience*, and *sheer motivation*. With respect to collocational frequency, there is value in considering not just the frequency of collocations in an external corpus, but also the collocates within collocations, since the results suggested that experts especially noticed collocations containing lower-frequency lemmas (and thus award higher ratings of lexical proficiency).

This suggestion that collocates held special prominence is supported by the rater comments. As noted, raters provided single-word and multiword lexical items, including collocations, to exemplify lexical sophistication. This finding, combined with the high number of times the term *collocation* was explicitly used, suggested that collocations were especially salient to examiners. Furthermore, in some cases, a specific collocate appeared to be particularly noticeable as some examiners gave the single word as an example and others gave the word as part of a collocation, for example *first-hand* versus *first-hand experience*, suggesting that it was collocate frequency, rather than collocation frequency, which drew attention. These cases exemplify how raters may differ in the extent to which multiword items are noticed, compared to single words, as well as the way in which sophistication is conceptualized.

With respect to the utility of the three-way classification of high-, mid-, low-frequency collocations, it was originally hypothesized that there would be a clear distinction between texts systematically varying based on these frequency categories. The results can be seen to generally support this view, especially that a "mid" category is informative, since high-mid and high-low contrasts were significant. However, the mid-low contrast was not significant, perhaps due to the coarse-grained frequency "buckets," and it may be that more fine-grained divisions at the lower frequencies

https://doi.org/10.1017/S0272263125000075 Published online by Cambridge University Press

would have been more revealing (see Kremmel, 2016), albeit at the expense of practical utility for practitioners who use online frequency profiling tools.

RQ2: Lack of significance of collocational accuracy

It was predicted that the high accuracy level would lead to higher LR ratings based on previous literature that found collocational accuracy to be an important aspect of judgements of proficiency (e.g., Laufer & Waldman, 2011; Nesselhauf, 2005). It was therefore contrary to expectation that no such effect was uncovered in the ratings. One potential explanation is that the quantity of collocation errors between the Low accuracy and High accuracy versions was insufficient. In other words, adding six additional collocation errors to a text of 250 words was too small a manipulation, regardless of the base CEFR level.

A second, and perhaps more likely, explanation relates to the level of *error gravity*, namely the impact of the errors on communication. In this study, the meaning remains clear in all the collocation errors, for example, *positive school.* This consistent "light" error gravity likely decreased the impact of the collocation inaccuracies. In addition, a word combination such as *positive school* may have been interpreted by raters as creative language use rather than "wrong," as from a phraseological perspective such combinations are not restricted collocations in the same way that *slim chance* or *make a mistake* are. This level of error gravity and type of inaccurate collocation also matches the IELTS lexical descriptors for band 6 and below, which focus on lexical accuracy in terms of impact on communication, for example, B6: "makes some errors in spelling and/or word formation, *but they do not impede communication*" (emphasis added). If the raters, as expected, closely followed the rubric descriptors, then this wording may also help to explain why accuracy as operationalized in this study did not emerge as a strong predictor of the ratings at the B1/B2 levels, though it does not explain the lack of significant interaction between B2-C1 and accuracy.

In contrast to the results of the linear regression model, the raters' comments demonstrated that lexical accuracy in its many forms was a feature they considered important. Collocations with inaccurate word choice were frequently noted, for example, *positive school*, *study at money*, and *straight contribution.* As a result, writers at all three CEFR levels were often described as "risk takers," that is, writers with higher sophistication but lower accuracy. These rater data further support the hypothesis that the lack of significance for collocational accuracy in this study can likely be attributed to experimental design.

Pedagogical implications: Choosing which collocations to teach

At present, collocation instruction is common in many contexts, but the selection of which collocations to teach often remains unprincipled (Macis & Schmitt, 2016). A general rule-of-thumb of any vocabulary selection is to consider the cost-benefit principle so that learners get the best return for the time invested in learning. Frequency is one way of deciding this benefit and has been traditionally used to determine text coverage (the number of known lemmas/word families needed to cover a certain percentage of texts) and to create frequency lists.

Some single-lemma frequency lists are widely used in general English (e.g., New General Service List [NGSL]; Browne et al., 2013) and English for Academic Purposes [EAP] (e.g., Academic Word List [AWL]; Coxhead, 2000). However, there are few

https://doi.org/10.1017/S0272263125000075 Published online by Cambridge University Press

widely used collocation frequency lists (though see Ackerman & Che, 2013; Durrant, 2009; Shin & Nation, 2007). A practical approach for teachers is to focus on formulaic sequences containing items from established frequency-based lists such as the AWL (as advocated by Coxhead [2020])—essentially a collocate frequency approach.

However, findings from studies such as this paper suggest that the inclusion of some lexis in the K10+ bands in learning goals is also worthwhile. Currently such lexis is not supported in frequency list approaches to vocabulary selection, either for individual words or collocations. But, as Vilkaitė-Lozdienė and Schmitt (2020, p. 88) caution, "frequency lists should be seen more as a useful indication rather than a prescription." While high-frequency words are crucial for comprehension and text coverage, the findings from this study suggest that knowledge of lower-frequency items is particularly important for productive skills, especially in assessment contexts.

Instead of *replacing* frequency-based lists/materials that focus on K1-2 lexis, one option is to *supplement* the existing curricula with judiciously selected K3-9 *and* K10+ lexical items. In doing so, vocabulary pedagogy practices can still be evidence-based rather than solely intuition-based, but more responsive to individuals' needs. For example, the following categories represent low-frequency collocations which nonetheless have wide academic generalizability and therefore high cost-benefit:

1. *Discourse markers:* By replacing or inserting K10+ words into discourse markers (e.g., *in my case* [K1] → *not to generalize* [K13]), students can apply these formulaic sequences in a range of academic text types to good effect, concurrently improving the sophistication of their lexis while demonstrating flexible use of cohesive devices.
2. *Synonyms for other K1-2 collocations*: Collocations containing nouns are the most frequent type of lexical collocation (Nizonkiza & Van de Poel, 2019) and are a key attribute of academic prose. However, learners tend to underuse noun forms in their own writing in favor of verbs (Naismith & Juffs, 2021). High-frequency collocations can therefore be naturally replaced with low-frequency collocations containing noun forms (e.g., *learn about* [K1] → *gain proficiency in* [K10]).
3. *Domain-specific, specialized lexis*: For many students, it is necessary to not only know general academic English vocabulary, but also lexis specific to their studies and careers (Coxhead, 2020; Nation, 2013), for example *smart shopper* (K1) → *savvy shopper* (K11) in marketing. This type of specialized vocabulary is one of the greatest challenges that learners report (Dang & Dang, 2021).

Assessment implications

In the public IELTS descriptors, relativistic terminology is frequently used to distinguish between bands. For example, sophistication descriptors include "attempts to use less common vocabulary" (B6), "uses less common lexical items" (B7), and "skillfully uses uncommon lexical items" (B8). What is unclear is whether "vocabulary" and "lexical items" are synonymous or intended to distinguish between single and multiword lexical items, or exactly what frequencies "less common" and "uncommon" refer to. Research has shown that teachers debate the meanings of terms in descriptors (Claire, 2001) and have difficulty interpreting/applying relativistic terminology (Smith, 2000). A compromise could therefore limit the number of different modifiers for describing lexical use and to gloss elsewhere what approximate frequency ranges these terms are intended to encompass, illustrated with examples.

https://doi.org/10.1017/S0272263125000075 Published online by Cambridge University Press

Table 8. Band 6 Lexical Resource descriptors

Original descriptor	Amended descriptor
Uses an adequate range of vocabulary for the task	Uses an adequate range of **words and multiword expressions** for the task
Attempts to use less common vocabulary but with some inaccuracy	Attempts to use less common **words and multiword expressions** but with some inaccuracy
Makes some errors in spelling and/or word formation, but they do not impede communication	Makes some errors in spelling, word formation, **and word choice (including collocations)**, but they do not impede communication

Most key lexical dimensions are included in the descriptors at nearly every IELTS band level. However, consideration of formulaic sequences, including collocations, is lacking: only bands 7 and 8 include the term "collocation." It is true that formulaic sequences are a component of "vocabulary" and "lexical items," but without being explicitly described and mentioned at all levels, there is a danger that raters overlook or undervalue the many types of lexical items. This oversight may occur even though, as this paper has demonstrated, collocational sophistication (at least in terms of collocate frequency) is a significant factor in rating and is therefore deserving of explicit recognition. The wording of rubrics is important, and more experienced raters use more rubric-generated vocabulary to describe decisions and ratings (Wolfe et al., 1998). It therefore seems that the current scales do not adequately address key elements of collocational proficiency.

To exemplify how the descriptors might be updated, Table 8 contains the original Band 6 Lexical Resource descriptor and a potential amended version. This updated descriptor takes into consideration the beliefs of the raters in this study and the research findings supporting the importance of collocational proficiency. In doing so, it is intended to more clearly highlight a key element of lexical proficiency which at present is not given sufficient weight.

One challenge of writing descriptors is balancing specificity and practicality as there is very limited space available. As such, minimal alterations have been made (emphasized in bold) to make salient that formulaic sequence use is part of the existing descriptors for range, sophistication, and accuracy. Here we suggest the term *multiword expression* over *formulaic sequence* because the former is currently, in our experience, more widely used in the teaching community and more immediately accessible.

Conclusions

This paper reported on an investigation of expert IELTS examiner ratings of texts which had been manipulated in terms of their collocational frequency (a dimension of collocational sophistication) and accuracy. The resulting data showed that the frequency of lexical items in general was impactful, especially when less-frequent words were part of salient collocations. In general, the high-, mid-, and low-frequency categories were an appropriate method for identifying different levels of lexical sophistication, though the division between mid- and low-frequency, as operationalized here, was not entirely clear cut. Furthermore, while collocational accuracy seemed to be noticeable to raters, it did not impact the statistical models.

The main contributions of this mixed methods study are threefold. From a methodological standpoint, the careful text selection and normalization provides a model for future research. By carefully normalizing text length and validating the results, student

https://doi.org/10.1017/S0272263125000075 Published online by Cambridge University Press

essays can be used as research instruments without needing to account for text length and topic/prompt effects. In addition, the use of MFRM models to obtain fair scores prior to further inferential analysis remains uncommon in this field of research but shows merit in terms of accounting for individual rater variability first, before carrying out linguistic analysis. By including qualitative data from expert raters, the quantitative data can be better interpreted and receive additional validation.

The second contribution of this study is to classroom pedagogy for the teaching of lexis. Historically, the teaching of formulaic sequences and collocations has been neglected (Wolter, 2020), and even though there has been a resurgence in this area, the decision of which collocations to teach is often left to "the whims of individual teachers" rather than based on empirical research (Hanks, 2013, p. 424). The results of this study suggested that for students to improve the quality of their written academic English, it is beneficial to judiciously include some *very* low-frequency lexis, *even if* learning such lexis is of lesser benefit to developing receptive skills.

A third contribution of this study is to inform potential assessment training and scale design practices. Given the importance of assessment literacy for raters in delivering reliable assessments, it is critical to provide teachers and examiners with training and tools which help clarify the key elements of learners' lexis. As such, it is recommended that formulaic sequences be an explicit component of all band descriptors, and that the relationship between frequency descriptors and frequency bands be clarified.

Many of the limitations of this study result from conscious decisions regarding its methodological design. The experimental nature of the study required controlling features such as text length, the frequency bands, and accuracy of specific lexical items. The trade-off for this degree of control is the authenticity of the texts, the use of only one writing prompt, and the use of only three base texts from different proficiency levels, all of which may have had unintended and unmeasured effects on the ratings. In addition, the exclusion of levels of error gravity as a factor somewhat limits the conclusions that can be drawn about the collocational accuracy findings.

Future research might therefore include partial replications of this study but with adjustments to the texts to increase the difference in quantity of collocation errors or error severity between the low and high accuracy text versions. A pilot study could also be carried out to ascertain whether potentially inaccurate collocations are experienced as such. Adjusting collocation sophistication using a collocation frequency approach or other operationalizations of sophistication would also be informative. The qualitative element of this research, the raters' comments, could also be further explored through interviews or surveys to acquire a more thorough understanding of the raters' thought processes and beliefs about lexis and assessment. Through projects such as the current study and others in a similar vein, it will be possible to better understand the relationship between text quality as it is realized through learners' use of lexis and the way it is perceived by expert raters of high-stakes tests.

Supplementary material. The supplementary material for this article can be found at http://doi.org/10.1017/S0272263125000075.

Acknowledgements. We express our gratitude to the participants for their time, to the reviewers for their helpful suggestions, and to Drs. Matthew Kanwit, Melinda Fricke, Na-Rae Han, and Ute Römer-Barron for their feedback on an earlier version of this work. All errors are, of course, our own. This work was supported by the Social Sciences and Humanities Research Council of Canada and a Duolingo Research Grant.

Competing interest. The author(s) declare none.

References

Ackermann, K., & Chen, Y.-H. (2013). Developing the Academic Collocation List (ACL) – A corpus-driven and expert-judged approach. *Journal of English for Academic Purposes*, *12*, 235–247. https://doi.org/10.1016/j.jeap.2013.08.002

Agustín Llach, M. d. P. (2011). *Lexical errors and accuracy in foreign language writing*. Channel View Publications.

Arnaud, P. J. (1984). The lexical richness of L2 written productions and the validity of vocabulary tests. In T. Culhane, C. Klein-Braley, & D. K. Stevenson (Eds.), *Practice and problems in language testing papers from the international symposium on language testing* (pp. 14–28). University of Essex.

Aryadoust, V., Ng, L. Y., & Sayama, H. (2021). A comprehensive review of Rasch measurement in language assessment: Recommendations and guidelines for research. *Language Testing*, *38*(1), 6–40. https://doi.org/10.1177/0265532220927487

Attali, Y. (2016). A comparison of newly-trained and experienced raters on a standardized writing assessment. *Language Testing*, *33*(1), 99–115. https://doi.org/10.1177/0265532215582283

Barkaoui, K. (2011). Think-aloud protocols in research on essay rating: An empirical study of their veridicality and reactivity. *Language Testing*, *28*(1), 51–75. https://doi.org/10.1177/0265532210376379

Bestgen, Y. (2017). Beyond single-word measures: L2 writing assessment, lexical richness and formulaic competence. *System*, *69*(1), 65–78. https://doi.org/10.1016/j.system.2017.08.004

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. https://doi.org/10.1016/j.jslw.2014.09.004

Brown, A. (2006). An examination of the rating process in the revised IELTS Speaking Test. In P. McGovern & S. Walsh (Eds.), *IELTS Research Reports 2006* (Vol. *6*, pp. 1–30). IELTS Australia.

Browne, C., Culligan, B., & Phillips, J. (2013). *The New General Service List*. http://www.newgeneralservicelist.org

Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, *16*(1), 22–29.

Claire, S. (2001). Assessment and moderation in CSWE: Processes, performances and tasks. In G. Brindley & C. Burrows (Eds.), *Studies in immigrant English language assessment Volume 2* (pp. 15–57). National Centre for English Language Teaching and Research and Macquarie University.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

Conklin, K., & Schmitt, N. (2012). The processing of formulaic language. *Annual Review of Applied Linguistics*, *32*, 45–61. https://doi.org/10.1017/S0267190512000074

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*(2), 213–238. https://doi.org/10.2307/3587951

Coxhead, A. (2020). Academic Vocabulary. In S. Webb (Ed.), *The Routledge Handbook of Vocabulary Studies* (pp. 97–110). Routledge. https://doi.org/10.4324/9780429291586-7

Creswell, J. W., & Plano Clark, V. L. (2011). *Designing and conducting mixed methods research*. SAGE Publications.

Crossley, S. A., & McNamara, D. S. (2012). Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication. *Journal of Research in Reading*, *35*(2), 115–135. https://doi.org/10.1111/j.1467-9817.2010.01449.x

Crossley, S. A., Salsbury, T., & McNamara, D. S. (2012). Predicting the proficiency level of language learners using lexical indices. *Language Testing*, *29*(2), 243–263. https://doi.org/10.1177/0265532211419331

Crossley, S. A., Salsbury, T., & Mcnamara, D. S. (2013). Validating lexical measures using human scores of lexical proficiency. In S. Jarvis & M. Daller (Eds.), *Vocabulary knowledge: Human ratings and automated measures* (pp. 105–134). John Benjamins. https://doi.org/10.1075/sibil.47.06ch4

Crossley, S. A., Salsbury, T., & Mcnamara, D. S. (2015). Assessing lexical proficiency using analytic ratings: A case for collocation accuracy. *Applied Linguistics*, *36*(5), 570–590. https://doi.org/10.1093/applin/amt056

Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing*, *7*(1), 31–51. https://doi.org/10.1177/026553229000700104

Daller, H., Turlik, J., & Weir, I. (2013). Vocabulary acquisition and the learning curve. In S. Jarvis & M. Daller (Eds.), *Vocabulary knowledge: Human ratings and automated measures* (pp. 185–218). John Benjamins. https://doi.org/10.1075/sibil.47.09ch7

Dang, C. N., & Dang, T. N. Y. (2021). The predictive validity of the IELTS test and contribution of IELTS preparation courses to international students' subsequent academic study: Insights from Vietnamese international students in the UK. *RELC Journal*. https://doi.org/10.1177/0033688220985533

Davies, M. (2008). *The Corpus of Contemporary American English (COCA)* [linguistic corpora]. https://www.english-corpora.org/coca/

Durrant, P. (2009). Investigating the viability of a collocation list for students of English for academic purposes. *English for Specific Purposes*, *28*(3), 157–169. https://doi.org/10.1016/j.esp.2009.02.002

Durrant, P. (2014). Corpus frequency and second language learners' knowledge of collocations: A meta-analysis. *International Journal of Corpus Linguistics*, *19*(4), 443–477. https://doi.org/10.1075/ijcl.19.4.01dur

Durrant, P. (2019). Formulaic language in English for Academic Purposes. In A. Siyanova-Chanturia & A. Pellicer-Sánchez (Eds.), *Understanding Formulaic Language: A Second Language Acquisition Perspective* (pp. 211–227). Routledge.

Durrant, P., & Schmitt, N. (2009). To what extent do native and non-native writers make use of collocations? *International Review of Applied Linguistics in Language Teaching*, *47*(2), 157–177. https://doi.org/10.1515/iral.2009.007

Ebrahimi, A. (2017). *Measuring productive depth of vocabulary knowledge of the most frequent words* (Publication Number 4894) Electronic Thesis and Dissertation Repository. https://ir.lib.uwo.ca/etd/4894

Eckes, T. (2012). Operational rater types in writing assessment: Linking rater cognition to rater behavior. *Language Assessment Quarterly*, *9*(3), 270–292. https://doi.org/10.1080/15434303.2011.649381

Egushi, M. and Kyle, K. (2020), Continuing to explore the multidimensional nature of lexical sophistication: The case of oral proficiency interviews. *The Modern Language Journal*, *104*, 381–400. https://doi.org/10.1111/modl.12637

Ellis, R. (2008). *The Study of Second Language Acquisition* (2nd ed.). Oxford University Press.

Evert, S. (2009). Corpora and collocations. *Corpus Linguistics: An International Handbook*. https://doi.org/10.1515/9783110213881.2.1212

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Fritz, E., & Ruegg, R. (2013). Rater sensitivity to lexical accuracy, sophistication and range when assessing writing. *Assessing Writing*, *18*(2), 173–181. https://doi.org/10.1016/j.asw.2013.02.001

Garner, J. (2022). The cross-sectional development of verb-noun collocations as constructions in L2 writing. *International Review of Applied Linguistics in Language Teaching*, *60*(3), 909–935. https://doi.org/10.1515/iral-2019-0169

Garner, J., Crossley, S. & Kyle, K. (2019). N-gram measures and L2 writing proficiency. *System*, *80*, 176–187. https://doi.org/10.1016/j.system.2018.12.001

Garner, J., Crossley, S. & Kyle, K. (2020). Beginning and intermediate L2 writer's use of N-grams: an association measures study. *International Review of Applied Linguistics in Language Teaching*, *58*(1), 51–74. https://doi.org/10.1515/iral-2017-0089

Goh, C. C. M., & Ang-Aw, H. T. (2018). Teacher-examiners' explicit and enacted beliefs about proficiency indicators in national oral assessments. In D. Xerri & P. Vella Briffa (Eds.), *Teacher Involvement in High-Stakes Language Testing* (pp. 197–216). Springer International Publishing. https://doi.org/10.1007/978-3-319-77177-9_11

González Fernández, B., & Schmitt, N. (2015). How much collocation knowledge do L2 learners have? The effects of frequency and amount of exposure. *International Journal of Applied Linguistics*, *166*, 94–126. https://doi.org/10.1075/itl.166.1.03fer

Granger, S. (1998). Prefabricated patterns in advanced EFL writing: Collocations and formulae. In A. P. Cowie (Ed.), *Phraseology: theory, analysis, and applications* (pp. 145–160). Clarendon Press.

Granger, S. (2003). Error-tagged learner corpora and CALL: a promising synergy. *CALICO Journal*, *20*(3), 465–480.

Granger, S., & Bestgen, Y. (2014). The use of collocations by intermediate vs. advanced non-native writers: A bigram-based study. *International Review of Applied Linguistics in Language Teaching*, *52*(3), 229–252. https://doi.org/10.1515/iral-2014-0011

Guo, L., Crossley, S. A., & McNamara, D. S. (2013). Predicting human judgments of essay quality in both integrated and independent second language writing samples: a comparison study. *Assessing Writing*, *18*(3), 218–238. https://doi.org/10.1016/j.asw.2013.05.002

Gyllstad, H., & Wolter, B. (2016). Collocational processing in light of the phraseological continuum model: Does semantic transparency matter? *Language Learning*, *66*(2), 296–323. https://doi.org/10.1111/lang.12143

Ha, M.-J. (2013). Corpus-based analysis of collocational errors. *International Journal of Digital Content Technology and its Applications*, *7*(11), 100–108.

Hall, C., & Sheyholislami, J. (2013). Using appraisal theory to understand rater values: an examination of rater comments on ESL test essays. *The Journal of Writing Assessment*, *6*(1).

Hamp-Lyons, L. (1990). Second language writing: assessment issues. In B. Kroll (Ed.), *Second Language Writing* (pp. 69–87). Cambridge University Press. https://doi.org/10.1017/CBO9781139524551.009

Hanks, P. (2013). *Lexical Analysis: Norms and Exploitations*. MIT Press.

Hasselgren, A. (1994). Lexical teddy bears and advanced learners: A study into the ways Norwegian students cope with English vocabulary. *International Journal of Applied Linguistics*, *4*(2), 237–258. https://doi.org/10.1111/j.1473-4192.1994.tb00065.x

Henriksen, B. (2013). Research on L2 learners' collocational competence and development – a progress report. In C. Bardel, C. Lindqvist, & B. Laufer (Eds.), *EUROSLA Monographs Series 2* (pp. 29–56). European Second Language Association.

Howarth, P. (1998). Phraseology and second language proficiency. *Applied Linguistics*, *19*(1), 24–44. https://doi.org/10.1093/applin/19.1.24

IELTS. (n.d.-a). *IELTS Academic - paper sample tests.* https://ielts.org/take-a-test/preparation-resources/sample-test-questions/academic-test

IELTS. (n.d.-b). *IELTS scoring in detail.* https://ielts.org/take-a-test/preparation-resources/understanding-your-score/ielts-scoring-in-detail

IELTS. (n.d.-c). *Understanding your score.* https://www.ielts.org/about-the-test/how-ielts-is-scored

Jarvis, S. (2013). Defining and measuring lexical diversity. In S. Jarvis & M. Daller (Eds.), *Vocabulary knowledge: Human ratings and automated measures* (pp. 13–43). John Benjamins.

Jiang, J. (2009). Designing pedagogic materials to improve awareness and productive use of L2 collocations. In A. Barfield & H. Gyllstad (Eds.), *Researching collocations in another language: multiple interpretations* (pp. 99–113). Palgrave Macmillan.

Jiang, J., Bi, P., Xie, N. & Liu, H. (2023). Phraseological complexity and low- and intermediate-level L2 learners' writing quality. *International Review of Applied Linguistics in Language Teaching*, *61*(3), 765–790. https://doi.org/10.1515/iral-2019-0147

Juffs, A., Han, N.-R., & Naismith, B. (2020). *The University of Pittsburgh English Language Corpus (PELIC)* [linguistic corpora]. https://doi.org/10.5281/zenodo.3991977

Kim, 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*(1), 120–141. https://doi.org/10.1111/modl.12447

Kojima, M., & Kaneta, T. (2022). L2 writing and its internal correlates: A meta-analysis. In E. H. Jeon & Y. In'nami (Eds.), *Understanding L2 proficiency: Theoretical and meta-analytic investigations* (pp. 109–158). John Benjamins Publishing Company. https://doi.org/10.1075/bpa.13.05koj

Kojima, M., In'nami, Y., & Kaneta, T. (2022). L2 writing and its external correlates: A meta-analysis. In E. H. Jeon & Y. In'nami (Eds.), *Understanding L2 proficiency: Theoretical and meta-analytic investigations* (pp. 159–211). John Benjamins Publishing Company. https://doi.org/10.1075/bpa.13.06koj

Kremmel, B. (2016), Word families and frequency bands in vocabulary tests: Challenging conventions. *TESOL Quarterly*, *50*, 976–987. https://doi.org/10.1002/tesq.329

Kyle, K. & Crossley, S.A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly*, *49*(4), 757–786. https://doi.org/10.1002/tesq.194

Kyle, K., Crossley, S. A., & Jarvis, S. (2020). Assessing the validity of lexical diversity indices using direct judgements. *Language Assessment Quarterly*, 1–17. https://doi.org/10.1080/15434303.2020.1844205

Laufer, B. (1989). What percentage of text lexis is essential for comprehension? In C. Laurén & M. Nordman (Eds.), *Special Language: From Human Thinking to Thinking Machines* (pp. 316–323). Multilingual Matters.

Laufer, B., & Ravenhorst-Kalovski, G. C. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language*, *22*, 15–30.

Laufer, B., & Waldman, T. (2011). Verb-noun collocations in second language writing: A corpus analysis of learners' English. *Language Learning*, *61*(2), 647–672. https://doi.org/10.1111/j.1467-9922.2010.00621.x

Lee, S. (2019). L1 transfer, proficiency, and the recognition of L2 verb-noun collocations: A perspective from three languages. *International Review of Applied Linguistics in Language Teaching*, *59*(2), 181–208. https://doi.org/10.1515/iral-2018-0220

Lenko-Szymanska, A. (2019). *Defining and assessing lexical proficiency*. Routledge.

Levshina, N. (2015). *How to do linguistics with R: Data exploration and statistical analysis*. John Benjamins. https://doi.org/10.1075/z.195

Lim, G. S. (2011). The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters. *Language Testing*, *28*(4), 543–560. https://doi.org/10.1177/0265532211406422

Linacre, J. M. (1989, 1994). *Many-facet Rasch measurement*. MESA Press.

Linacre, J. M. (2020). *Facets computer program for many-facet Rasch measurement*. In (Version 3.83.4) [Computer software]. https://www.winsteps.com

Linnarud, M. (1986). *Lexis in composition: a performance analysis of Swedish learners' written English*. Liber Förlag.

Lu, X., & Hu, R. (2022). Sense-aware lexical sophistication indices and their relationship to second language writing quality. *Behavior Research Methods*, *54*, 1444–1460. https://doi.org/10.3758/s13428-021-01675-6

Lumley, T. (2005). *Assessing Second Language Writing*. Peter Lang.

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: implications for training. *Language Testing*, *12*(1), 54–71. https://doi.org/10.1177/026553229501200104

Lundell, F. F., & Lindqvist, C. (2012). Vocabulary aspects of advanced L2 French: Do lexical formulaic sequences and lexical richness develop at the same rate? *Language, Interaction and Acquisition*, *3*(1), 73–92. https://doi.org/10.1075/lia.3.1.05for

Macis, M., & Schmitt, N. (2016). Not just 'small potatoes': Knowledge of the idiomatic meanings of collocations. *Language Teaching Research*, *21*(3), 321–340. https://doi.org/10.1177/1362168816645957

McNamara, T. F., Knoch, U., & Fan, J. (2019). *Fairness, Justice and Language Assessment*. Oxford University Press.

Millar, N. (2011). The processing of malformed formulaic language. *Applied Linguistics*, *32*(2), 129–148, https://doi.org/10.1093/applin/amq035

Monteiro, K. R., Crossley, S. A., & Kyle, K. (2020). In search of new benchmarks: Using L2 lexical frequency and contextual diversity indices to assess second language writing. *Applied Linguistics*, *41*(2), 280–300. https://doi.org/10.1093/applin/amy056

Naismith, B., Han, N.-R., & Juffs, A. (2022). The University of Pittsburgh English Language Institute Corpus (PELIC). *International Journal of Learner Corpus Research*, *8*(1), 121–138. https://doi.org/10.1075/ijlcr.21002.nai

Naismith, B., & Juffs, A. (2021). Finding the sweet spot: Learners' productive knowledge of mid-frequency lexical items. *Language Teaching Research*. https://doi.org/10.1177/13621688211020412

Nation, I. S. P. (2013). *Learning vocabulary in another language* (2nd ed.). Cambridge University Press.

Nation, I. S. P., & Anthony, L. (2013). Mid-frequency readers. *Journal of Extensive Reading*, *1*, 5–16.

Nattinger, J. R., & DeCarrico, J. S. (1992). *Lexical phrases and language teaching*. Oxford University Press.

Nesselhauf, N. (2005). *Collocations in a learner corpus* (Vol. 14). John Benjamins Amsterdam.

Nizonkiza, D., & Van de Poel, K. (2019). Mind the gap: Towards determining which collocations to teach. *Stellenbosch Papers in Linguistics Plus (SPiL Plus)*, *56*, 13–30.

Öksüz, D., Brezina, V., & Rebuschat, P. (2021). Collocational processing in L1 and L2: The effects of word frequency, collocational frequency, and association. *Language Learning*, *71*(1), 55–98. https://doi.org/10.1111/lang.12427

Paquot, M. (2018). Phraseological competence: A missing component in university entrance language tests? Insights from a study of EFL learners' use of statistical collocations. *Language Assessment Quarterly*, *15*(1), 29–43. https://doi.org/10.1080/15434303.2017.1405421

Polio, C., & Shea, M. C. (2014). An investigation into current measures of linguistic accuracy in second language writing research. *Journal of Second Language Writing*, *26*, 10–27. https://doi.org/10.1016/j.jslw.2014.09.003

Polio, C. G. (1997). Measures of linguistic accuracy in second language writing research. *Language Learning*, *47*(1), 101–143. https://doi.org/10.1111/0023-8333.31997003

Qian, D. D., & Lin, L. H. F. (2020). The relationship between vocabulary knowledge and language proficiency. In S. Webb (Ed.), *The Routledge Handbook of Vocabulary Studies* (pp. 66–80). Routledge. https://doi.org/10.4324/9780429291586-5

R Development Core Team. (2019). *R: A language environment for statistical computing*. In R Foundation for Statistical Computing. https://www.R-project.org/

Read, J. (2000). *Assessing vocabulary*. Cambridge University Press. https://doi.org/10.1017/CBO9780511732942

Read, J. (2004). Plumbing the depths: How should the construct of vocabulary knowledge be defined? In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a second language: Selection, acquisition, and testing* (Vol. 10, pp. 209–227). John Benjamins.

Read, J., & Nation, I. S. P. (2006). An investigation of the lexical dimension of the IELTS Speaking Test. *IELTS Research Reports*, *6*, 207–231.

Santos, T. (1988). Professors' reactions to the academic writing of nonnative-speaking students. *TESOL Quarterly*, *22*(1), 69–90. https://doi.org/10.2307/3587062

Shin, D., & Nation, P. (2007). Beyond single words: the most frequent collocations in spoken English. *ELT Journal*, *62*(4), 339–348. https://doi.org/10.1093/elt/ccm091

Simpson-Vlach, R., & Ellis, N. C. (2010). An academic formulas list: New methods in phraseology research. *Applied Linguistics*, *31*(4), 487–512. https://doi.org/10.1093/applin/amp058

Siyanova-Chanturia, A., & Martinez, R. (2015). The idiom principle revisited. *Applied Linguistics*, *36*(5), 549–569. https://doi.org/10.1093/applin/amt054

Siyanova-Chanturia, A., & Pellicer-Sánchez, A. (2020). Formulaic language: Setting the scene. In A. Siyanova-Chanturia & A. Pellicer-Sánchez (Eds.), *Understanding Formulaic Language: A Second Language Acquisition Perspective* (pp. 1–16). Routledge. https://doi.org/10.4324/9781315206615-3

Siyanova-Chanturia, A., & Sidtis, D. V. L. (2019). What online processing tells us about formulaic language. In A. Siyanova-Chanturia & A. Pellicer-Sánchez (Eds.), *Understanding Formulaic Language: A Second Language Acquisition Perspective* (pp. 38–61). Routledge. https://doi.org/10.4324/9781315206615-3

Siyanova-Chanturia, A., & Spina, S. (2020). multiword expressions in second language writing: a large-scale longitudinal learner corpus study. *Language Learning*, *70*(2), 420–463. https://doi.org/10.1111/lang.12383

Smith, D. (2000). Rater judgements in the direct assessment of competency-based second language writing ability. In G. Brindley (Ed.), *Studies in immigrant English language assessment Volume 1* (pp. 159–189). National Centre for English Language Teaching and Research and Macquarie University.

Szudarski, Pawet. (2023). *Collocations, corpora and language learning*. Cambridge University Press.

Tsai, K.-J. (2015). Profiling the collocation use in ELT textbooks and learner writing. *Language Teaching Research*, *19*(6), 723–740. https://doi.org/10.1177/1362168814559801

Valdez, A., & Kaplan, C. D. (1998). Reducing selection bias in the use of focus groups to investigate hidden populations: the case of Mexican-American gang members from south Texas. *Drugs & Society*, *14*(1-2), 209–224. https://doi.org/10.1300/J023v14n01_15

Vilkaitė-Lozdienė, L., & Schmitt, N. (2020). Frequency as a guide for vocabulary usefulness. In S. Webb (Ed.), *The Routledge Handbook of Vocabulary Studies* (pp. 81–96). Routledge. https://doi.org/10.4324/9780429291586-6

Vögelin, C., Jansen, T., Keller, S. D., Machts, N., & Möller, J. (2019). The influence of lexical features on teacher judgements of ESL argumentative essays. *Assessing Writing*, *39*, 50–63. https://doi.org/10.1016/j.asw.2018.12.003

Wanner, L., Ramos, M. A., Vincze, O., Nazar, R., Ferraro, G., Mosqueira, E., & Prieto, S. (2013). Annotation of collocations in a learner corpus for building a learning environment. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *Twenty years of learner corpus research. Looking back, moving ahead* (pp. 493–503). Presses Universitaires de Louvain.

Wolfe, E. W., Kao, C.-W., & Ranney, M. (1998). Cognitive differences in proficient and nonproficient essay scorers. *Written Communication*, *15*(4), 465–492. https://doi.org/10.1177/0741088398015004002

Wolter, B. (2020). Key issues in teaching multiword items. In S. Webb (Ed.), *The Routledge Handbook of Vocabulary Studies* (pp. 493–510). Routledge. https://doi.org/10.4324/9780429291586-31

Wolter, B., & Yamashita, J. (2018). Word frequency, collocational frequency, L1 congruency, and proficiency in L2 collocational processing: What accounts for L2 performance? *Studies in Second Language Acquisition*, *40*(2), 395–416. https://doi.org/10.1017/S0272263117000237

Yoon, H.-J. (2016). Association strength of verb-noun combinations in experienced NS and less experienced NNS writing: Longitudinal and cross-sectional findings. *Journal of Second Language Writing*, *34*, 42–57. https://doi.org/10.1016/j.jslw.2016.11.001

Cite this article: Naismith, B., & Juffs, A. (2025). The impact of collocational proficiency features on expert ratings of L2 English learners' writing. *Studies in Second Language Acquisition,* 1–25. https://doi.org/10.1017/S0272263125000075